بسم الله الرحمن الرحیم

دانشکده مهندسی کامپیوتر

رساله دکتری مهندسی هوش مصنوعی

مکان‌یابی متون فارسی در تصاویر صحنه طبیعی با استفاده از روش جستجوی فرااکتشافی

نگارنده: جلیل قویدل نیچران

استاد راهنما

دکتر علیرضا احمدی‌فرد

استاد مشاور

دکتر مرتضی زاهدی

بهمن ۱۳۹۸

مدیریت تحصیلات تکمیلی

باسمه تعالی

شماره:

تاریخ: ۹۹/۲/۳

ویرایش:

فرم شماره 11: صورت جلسه نهایی دفاع از رساله دکتری (Ph.D)

بدینوسیله گواهی می شود آقای جلیل قویدل نیجران دانشجوی دکتری رشته مهندسی کامپیوتر-هوش مصنوعی به شماره دانشجویی ۹۱۲۵۲۵۵ ورودی ۹۱ در تاریخ ۹۸/۱۱/۲۹ از رساله نظری☐/ عملی☐ خود با عنوان **مکان‌یابی متون فارسی در تصاویر صحنه طبیعی با استفاده از روش جستجوی فراکتشافی**

دفاع و با اخذ نمره .... به درجه .... نائل گردید.

الف) درجه عالی: نمره 20-19 ☐	ب) درجه خیلی خوب: نمره 18/99 – 17 ☑
ج) درجه خوب: نمره 16/99 – 15 ☐	د) مردود: کمتر از 15 ☐

ردیف	هیئت داوران	نام و نام خانوادگی	مرتبه علمی	امضاء
۱	دکتر علیرضا احمدی‌فرد	استاد راهنما	دانشیار	
۲	دکتر مرتضی زاهدی	مشاور	استادیار	
۳	دکتر فرزین یغمائی	استاد مدعو خارجی	دانشیار	
۴	دکتر حمید حسن‌پور	استاد مدعو داخلی	استاد	
۵	دکتر منصور فاتح	استاد مدعو داخلی	استادیار	
۶	دکتر اسماعیل طحانیان	سرپرست (نماینده) تحصیلات تکمیلی دانشکده	استادیار	

مدیر محترم تحصیلات تکمیلی دانشگاه:

ضمن تأیید مراتب فوق مقرر فرمائید اقدامات لازم در خصوص انجام مراحل دانش آموختگی آقای جلیل قویدل نیجران بعمل آید.

نام و نام خانوادگی رئیس دانشکده:

تاریخ و امضاء و مهر دانشکده:

تقدیم به همسر مهربانم...

سپاس‌گزاری...

از استاد بزرگوارم جناب آقای دکتر احمدی‌فرد برای تمام حمایت‌ها و زحمات بی دریغ‌شان سپاسگزاری می‌کنم.

جلیل قویدل نیچران

بهمن ۱۳۹۸

تعهد نامه

اینجانب **جلیل قویدل نیچران** دانشجوی دکتری رشته **مهندسی هوش مصنوعی** دانشکده **مهندسی کامپیوتر** دانشگاه صنعتی شاهرود، نویسنده رساله با عنوان **مکان‌یابی متون فارسی در تصاویر صحنه طبیعی با استفاده از روش جستجوی فرااکتشافی** ، تحت راهنمایی دکتر **علیرضا احمدی‌فرد** متعهد می‌شوم:

- تحقیقات در این رساله توسط اینجانب انجام شده است و از صحت و اصالت برخوردار است.
- در استفاده از نتایج پژوهش‌های دیگر پژوهش‌گران، به مرجع مورد استفاده استناد شده است.
- مطالب این رساله، تا کنون توسط خود، یا فرد دیگری برای دریافت هیچ نوع مدرک یا امتیازی در هیچجا ارایه نشده است.
- حقوق معنوی این اثر، به دانشگاه صنعتی شاهرود تعلق دارد، و مقالات مستخرج با نام ‘‘ دانشگاه صنعتی شاهرود ‘‘ یا ‘‘ Shahrood University of Technology ‘‘ به چاپ خواهد رسید.
- حقوق معنوی تمام افرادی که در به‌دست آوردن نتایج اصلی رساله تاثیرگذار بوده‌اند، در مقالات مستخرج از رساله رعایت می‌گردد.
- در تمام مراحل انجام این رساله، در مواردی که از موجود زنده (یا بافت‌های آنها) استفاده شده است، ضوابط و اصول اخلاقی رعایت شده است.
- در تمام مراحل انجام این رساله، در مواردی که به حوزه اطلاعات شخصی افراد دسترسی یافته (یا استفاده شده است)، اصل رازداری و اصول اخلاق انسانی رعایت شده است.

جلیل قویدل نیچران

بهمن ۱۳۹۸

مالکیت نتایج و حق نشر

- تمام حقوق معنوی این اثر و محصولات آن (مقالات مستخرج، کتاب، برنامه‌های رایانه‌ای، نرم‌افزارها و تجهیزات ساخته شده) متعلق به دانشگاه صنعتی شاهرود می‌باشد. این مطلب باید به نحو مقتضی، در تولیدات علمی مربوطه ذکر شود.
- استفاده از اطلاعات و نتایج موجود در این رساله بدون ذکر منبع مجاز نمی‌باشد.

چکیده

هدف از این رساله مکان‌یابی متون فارسی در تصاویر صحنه طبیعی است. مکان‌یابی متن، عبارت است از پیدا کردن و محدوده‌گذاری تمامی نواحی در یک تصویر که انسان آن‌ها را متن در نظر می‌گیرد. به دلیل متغیر بودن اندازه، فونت و رنگ نوشته‌ها و وجود بافت در پس‌زمینه، این مسئله پرچالش بوده و باز به شمار می‌رود. چالش‌های اشاره شده برای زبان فارسی که الفبای آن در بیشتر موارد به هم چسبیده است بیشتر می شود؛ چسبیده بودن نویسه‌های الفبای فارسی، باعث ایجاد شباهت زیاد بین کلمات فارسی و شاخ و برگ درختان و موارد مشابه می شود.

در این رساله چهار روش جهت تشخیص متن فارسی و لاتین در تصاویر صحنه‌های طبیعی ارائه می گردد؛ هر روش جهت برطرف کردن مشکلات روش پیشین ارائه گردیده است. روش اول با نام «امضای رنگ لبه‌ها» مفهومی به نام لبه‌های رنگی را معرفی کرده و با استفاده از آن به مکان‌یابی متون می‌پردازد. این روش پس از استخراج لبه‌های تصویر ورودی، با استفاده از الگوریتم انتقال میانگین، رنگ‌های اطراف لبه‌ها را استخراج کرده و از این رنگ‌ها برای تولید نواحی کاندیدای متنی استفاده می‌کند. این نواحی کاندید در ادامه با معرفی چندین ویژگی جدید مبتنی بر لبه و یادگیری دیکشنری به کمک یک کلاسه‌بند آبشاری کلاسه‌بندی می‌شوند. در روش دوم با نام «تبدیل رنگ لبه‌ها»، لبه‌های استخراج شده از تصویر بهبود داده می‌شوند. همچنین عملگری برای به دست آوردن رنگ‌های اطراف لبه‌ها معرفی می‌شود که سریع‌تر و دقیق‌تر از روش پیشین است.

روش سوم، «تبدیل رنگ عمیق»، الگوریتم ایجاد نواحی کاندیدای روش دوم را بهبود داده و امکان مکان‌یابی متون با پس‌زمینه غیرهموار را فراهم می‌آورد. همچنین در این روش یک شبکه عصبی کانولوشنی با ساختاری جدید برای جداسازی مناطق متنی و غیرمتنی معرفی می‌شود. روش چهارم با نام «شبکه فرااکتشافی» از شبکه عصبی روش سوم به صورت مستقل برای مکان‌یابی متون استفاده می‌کند. خروجی این شبکه که درصد متن داخل ناحیه کاندیدا را نشان می‌دهد در یک الگوریتم فرااکتشافی جدید مورد استفاده قرار می‌گیرد. این روش با پخش کردن چندین پنجره در بخش‌های مختلف تصویر و استفاده از شبکه کانولوشنی به صورت تدریجی پنجره‌ها را به سوی بخش‌های متنی سوق می‌دهد.

روش‌های پیشنهادی با استفاده از مجموعه تصاویر فارسی «فارست» و همچنین مجموعه تصاویر شناخته شده $ICDAR2013$ برای تصاویر انگلیسی با همدیگر و با چندین روش شناخته شده دیگر مورد مقایسه قرار گرفته‌اند. تبدیل رنگ عمیق با مقدار f-measure برابر با ۸۶/۶۴ برای تصاویر انگلیسی و ۹۱/۵۸ برای تصاویر فارسی بهتر از روش‌های دیگر عمل کرده است.

کلمات کلیدی: مکان‌یابی متون فارسی، تصاویر صحنه طبیعی، تبدیل رنگ لبه‌ها، الگوریتم رشد لبه، یادگیری عمیق، روش جستجوی فرااکتشافی

لیست مقالات مستخرج از رساله

1. J. Ghavidel and A. Ahmadyfard, “Edge color transform: a new operator for natural scene text localization,” Multimed Tools Appl, vol. 77, no. 6, pp. 7615–7636
2. J. Ghavidel, A. Ahmadyfard, and M. Zahedi, “Natural scene text localization using edge color signature,” International Journal of Nonlinear Analysis and Applications, vol. 10, no. 1, pp. 229–237

فهرست مطالب

فهرست تصاویر ق

فهرست جداول ث

۱ مقدمه ۱

۱.۱ مقدمه و کاربردها ۱

۲.۱ تعریف مسئله و چالش‌های آن ۳

۳.۱ اهداف و پیش‌فرض‌ها ۴

۴.۱ نوآوری‌های پژوهش ۵

۵.۱ ساختار رساله ۷

۲ مروری بر کارهای گذشته ۹

۱.۲ مقدمه ۹

۲.۲ روش‌های مبتنی بر پنجره لغزان ۱۰

۳.۲ روش‌های مبتنی بر آنالیز مؤلفه‌های متصل ۱۲

۳ یادگیری عمیق ۲۱

۱.۳ مقدمه ۲۱

۲.۳ مزایای استفاده از شبکه‌های کانولوشنی ۲۲

۳.۳ لایه کانولوشن ۲۳

۴.۳ لایه‌ی ادغام بیشینه ۲۶
۵.۳ لایه کاملاً متصل ۲۷
۶.۳ توابع فعالیت ۲۷
۷.۳ ساختار کلی شبکه‌های کانولوشنی ۲۹
۸.۳ جمع‌بندی ۳۱

۴ امضای رنگ لبه‌ها ۳۳

۱.۴ مقدمه ۳۳
۲.۴ ایجاد هرمی از مقیاس‌های مختلف تصویر ورودی ۳۴
۳.۴ استخراج لبه‌ها ۳۶
۴.۴ هرس کردن پیکسل‌های غیر متنی ۳۷
۵.۴ امضای رنگ لبه‌ها ۴۰
۱.۵.۴ الگوریتم انتقال میانگین ۴۰
۶.۴ خوشه‌بندی نقاط لبه با استفاده از امضای رنگ ۴۳
۷.۴ استخراج ناحیه ۴۵
۸.۴ کلاسه‌بندی ۴۶
۱.۸.۴ سه لایه ابتدایی ۴۷
۲.۸.۴ لایه چهارم ۴۹
۳.۸.۴ لایه پنجم ۵۱
۴.۸.۴ تولید مناطق متنی ۵۳
۹.۴ حذف مناطق شبه متنی و ادغام مناطق هم‌پوشان ۵۵
۱۰.۴ نتایج تجربی ۵۶
۱۱.۴ جمع‌بندی ۶۳

۵ تبدیل رنگ لبه‌ها ۶۵

۱.۵ مقدمه ۶۵

۲.۵ به دست آوردن لبه ها ۶۶

۳.۵ تبدیل رنگ لبه‌ها ۶۶

۴.۵ الگوریتم رشد لبه ۶۹

۵.۵ نتایج به دست آمده ۷۳

۶.۵ جمع‌بندی ۷۳

۶ تبدیل رنگ عمیق و شبکه فرا اکتشافی ۷۷

۱.۶ مقدمه ۷۷

۲.۶ بهبود الگوریتم فصل گذشته ۷۸

۳.۶ شبکه عصبی کانولوشنی برای تشخیص نواحی متنی ۸۰

۴.۶ نتایج تجربی ۸۵

۵.۶ روش مبتنی بر یادگیری عمیق و الگوریتم فرا اکتشافی ۸۶

۶.۶ نتایج تجربی ۹۰

۷.۶ جمع‌بندی ۹۲

۷ نتیجه‌گیری و کارهای آینده ۹۵

۱.۷ مقدمه ۹۵

۲.۷ مقایسه روش‌های پیشنهادی ۹۶

۳.۷ مواردی که متن به درستی تشخیص داده نمی‌شود ۹۷

۴.۷ مواردی از تشخیص اشتباه مناطق غیر متنی ۹۹

۵.۷ پیشنهاد برای کارهای آینده ۱۰۲

مراجع ۱۰۵

فهرست تصاویر

۱.۱ مراحل یک سیستم استخراج متن از تصویر ۳

۱.۳ یادگیری ساختار سلسله مراتبی توسط شبکه‌های کانولوشنی [۱] ۲۴
۲.۳ نحوه ساخت نقشه‌های ویژگی در شبکه‌های کانولوشنی با استفاده از فیلترها که به صورت خطچین نشان داده شده‌اند. ۲۵
۳.۳ نحوه کارکرد لایه ادغام بیشینه ۲۷
۴.۳ لایه‌های کاملاً متصل ۲۸
۵.۳ مشخصات شبکه $VGG16$: برای لایه‌های کانولوشنی اندازه فیلترها و تعداد آن‌ها ذکر شده است. برای لایه‌های کاملاً‌ـ متصل نیز تعداد نرون‌ها برای هر لایه ذکر شده است. ۲۹

۱.۴ فلوچارت روش پیشنهادی اول ۳۵
۲.۴ تاثیر کاهش ابعاد تصویر در کیفیت لبه‌های خروجی الگوریتم ۳۶
۳.۴ مقایسه تعداد پیکسل‌های لبه استخراج شده توسط روش‌های سوبل و کنی برای یک تصویر نمونه آ) تصویر ورودی ب) خروجی روش کنی ج) خروجی روش سوبل ۳۷
۴.۴ مقایسه عملکرد روش‌های سوبل و کنی در مناطق متنی (لبه‌ها با رنگ مشکی نشان داده شده‌اند) آ) تصویر ورودی ب) برشی از تصویر ورودی ج) خروجی روش کنی د) خروجی روش سوبل ۳۸

۵.۴ تاثیر استفاده از قواعد هندسی برای هرس کردن نواحی غیر متنی، آ) تصویر ورودی ب) لبه‌های تصویر ورودی ج) تاثیر اعمال محدودیت (۱.۴) د) تاثیر اعمال محدویت (۲.۴) ه) تاثیر اعمال محدودیت (۳.۴) و) تاثیر اعمال محدودیت (۴.۴) ۴۱

۶.۴ نحوه بررسی کاندیداهای مختلف برای به دست آوردن مستطیل محیط کمینه ۴۵

۷.۴ نحوه استخراج ناحیه کاندید از تصویر آ) یافتن زاویه ناحیه با افق ب) ناحیه استخراج شده ۴۶

۸.۴ جزئیات ویژگی‌های مورد استفاده در لایه‌های مختلف کلاسه‌بند آبشاری ۴۸

۹.۴ نمونه‌هایی از تصاویر متنی تصنعی تولید شده آ) تصاویر تصنعی انگلیسی ب) تصاویر تصنعی انگلیسی ۵۵

۱۰.۴ نمونه‌هایی از تصاویر موجود در مجموعه داده ICDAR 2013 . . . . ۵۷

۱۱.۴ نمونه‌هایی از تصاویر موجود در مجموعه داده فارست ۵۸

۱۲.۴ برخی از خروجی‌های روش پیشنهادی برای مجموعه داده فارست . . . ۶۱

۱۳.۴ برخی از خروجی‌های روش پیشنهادی برای مجموعه داده فارست . . . ۶۲

۱.۵ تاثیر ترکیب خروجی لبه‌های به دست آمده از سه کانال رنگی قرمز، سبز و آبی روی کیفیت لبه‌ها آ) تصویر ورودی ب) کانال رنگی قرمز ج) کانال رنگی سبز د) کانال رنگی آبی ه) لبه‌های کانال قرمز و) لبه‌های کانال سبز ز) لبه‌های کانال آبی ح) لبه‌های تصویر خاکستری ط) لبه‌های حاصل از ترکیب سه کانال رنگی ۶۷

۲.۵ نحوه عملکرد تبدیل رنگ لبه‌ها آ) تصویر ورودی که در آن لبه‌ها با رنگ سفید مشخص شده‌اند ب) نحوه به دست آوردن امضای رنگ لبه‌ها ج)یکی از خروجی‌های عملگر تبدیل رنگ لبه‌ها ۶۸

۳.۵ یک همسایگی با مقدار $\delta = ۱۵$ که مرکز همسایگی با علامت + نشان داده شده است ۶۹

۴.۵ نواحی استخراج شده توسط الگوریتم رشد لبه برای یک تصویر نمونه . . ۷۱
۵.۵ الگوریتم رشد لبه ۷۲
۶.۵ برخی از خروجی‌های روش پیشنهادی برای مجموعه داده $ICDAR2013$ ۷۴
۷.۵ برخی از خروجی‌های روش پیشنهادی برای مجموعه داده فارست . . . ۷۵

۱.۶ خروجی تبدیل رنگ لبه‌ها و الگوریتم رشد لبه برای یک تصویر با پس‌زمینه پیچیده (برای بهبود کیفیت، کنتراست تصاویر افزایش داده شده است) آ) تصویر ورودی ب) لبه‌های تصویر ورودی ج_د) دو تصویر لبه رنگی ه_و) خروجی الگوریتم برای بخش‌های متنی ۷۸
۲.۶ اثر پس‌زمینه غیریکنواخت در خروجی الگوریتم رشد لبه آ) برشی از بخش متنی یک تصویر با پس‌زمینه غیر یکنواخت ب_ج) خروجی‌های تبدیل رنگ لبه‌ها د_ه) بخش‌های متنی جدا شده در هر بخش و) نتیجه ترکیب نواحی موجود در تصاویر بخش‌های «د» و «ه» ۷۹
۳.۶ جزئیات لایه ادغام ناحیه مطلوب برای یک کانال ۸۱
۴.۶ جزئیات شبکه $FastRCNN$ [۲] ۸۲
۵.۶ معماری شبکه $VGG16$ ۸۳
۶.۶ معماری شبکه پیشنهادی ۸۴
۷.۶ برخی از خروجی‌های روش پیشنهادی برای مجموعه داده $ICDAR2013$ ۸۷
۸.۶ برخی از خروجی‌های روش پیشنهادی برای مجموعه داده فارست . . . ۸۸
۹.۶ روند تدریجی همگرایی به بخش متنی برای یک تصویر نمونه ۸۹
۱۰.۶ عملیات ممکن برای هر سلول ۹۰
۱۱.۶ خروجی الگوریتم تپه‌نوردی برای دو تصویر نمونه ۹۱
۱۲.۶ برخی از خروجی‌های روش پیشنهادی مبتنی بر تپه‌نوردی برای مجموعه داده $ICDAR2013$ ۹۳
۱۳.۶ برخی از خروجی‌های روش پیشنهادی مبتنی بر الگوریتم تپه‌نوردی برای مجموعه داده فارست ۹۴

۱.۷ برخی از مواردی که روش امضای رنگ لبه‌ها قادر به تشخیص متن نیست ۹۸
۲.۷ برخی از مواردی که روش تبدیل رنگ لبه‌ها قادر به تشخیص متن نیست ۹۸
۳.۷ برخی از مواردی که روش شبکه فرا اکتشافی قادر به تشخیص متن نیست. بخش‌های متنی تشخیص داده نشده برای مشاهده بهتر در بالای هر تصویر به صورت بریده شده نشان داده شده‌اند ۹۹
۴.۷ برخی از تصاویری که هیچ یک از روش‌های پیشنهادی قادر به تشخیص متن نبوده‌اند. ردیف اول برای متون با یک کاراکتر و ردیف دوم برای تصاویر با کنتراست بسیار کم ۱۰۰
۵.۷ برخی از نواحی که به دلیل شباهت به متن به اشتباه متن تشخیص داده شده‌اند ۱۰۱

فهرست جداول

۱.۴ تأثیر اضافه کردن مجموعه تصاویر متنی تصنعی بر کارایی سیستم پیشنهادی ۶۰
۲.۴ نتایج روش پیشنهادی روی مجموعه داده ICDAR 2013 . . . . ۶۳
۳.۴ نتایج روش پیشنهادی روی مجموعه تصاویر فارست ۶۳

۱.۵ نتایج روش پیشنهادی روی مجموعه داده ICDAR 2013 . . . . ۷۳
۲.۵ نتایج روش پیشنهادی روی مجموعه تصاویر فارست ۷۳

۱.۶ نتایج روش پیشنهادی روی مجموعه داده ICDAR 2013 . . . . ۸۶
۲.۶ نتایج روش پیشنهادی روی مجموعه تصاویر فارست ۸۶
۳.۶ نتایج روش پیشنهادی روی مجموعه داده ICDAR 2013 . . . . ۹۱
۴.۶ نتایج روش پیشنهادی روی مجموعه تصاویر فارست ۹۱

۱.۷ مقایسه روش‌های پیشنهادی روی مجموعه داده ICDAR 2013 . . . . . ۹۶
۲.۷ مقایسه روش‌های پیشنهادی روی مجموعه داده فارست ۹۷
۳.۷ مقایسه روش‌های پیشنهادی با همدیگر و نقاط و ضعف و قوت آن‌ها . . ۱۰۳

فصل ۱

مقدمه

۱.۱ مقدمه و کاربردها

امروزه دوربین‌های دستی و همچنین دستگاه‌هایی مانند تلفن‌های همراه هوشمند و تبلت‌ها که در آن‌ها دوربین تعبیه شده است بسیار همه‌گیر شده‌اند و تجزیه و تحلیل تصاویر دیجیتال نیز در بینایی ماشین بیش از پیش مورد توجه قرار گرفته است. انسان‌ها به طور روزمره به مناسبت‌ها و دلایل مختلف مانند سفر، مسابقات ورزشی و گردش و موارد مشابه در حال تصویربرداری از محیط اطراف خود هستند و متون موجود در تصاویر یکی از اولین و ساده‌ترین راه‌ها برای درک تصاویر و فهم محتویات آن‌ها هستند. استخراج متن[۱] از تصاویر می‌تواند کمک شایانی به درک محیطی که از آن تصویربرداری صورت گرفته است بنماید. سیستم استخراج متن می‌تواند به عنوان بخشی از یک سیستم بزرگ‌تر، برای ارتباط بهتر محیط و انسان، مورد استفاده قرار گیرد که این ارتباط را یک واسط برقرار می‌کند؛ به عبارت

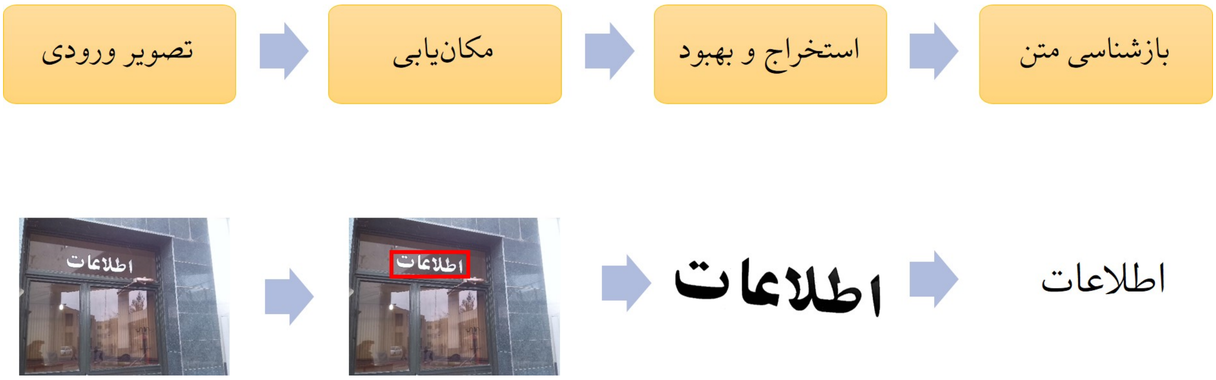
---
[۱] Text Extraction

دیگر، این سیستم می‌تواند بخشی از یک سیستم بزرگ‌تر مانند یک گوشی همراه یا اتومبیل باشد و با تصویربرداری از محیط و تشخیص وجود متن در صحنه‌ها و تعیین دقیق محل متن و سپس جدا کردن متن از زمینه و شناسایی آن، درک متن موجود در صحنه‌ها را ممکن سازد.
سیستم‌های استخراج متن تصویر می‌توانند کاربردهای زیادی داشته باشند که در ادامه به چند مورد از آن‌ها اشاره شده است.

- **کمک به افراد دچار مشکل بینایی:** افرادی که با مشکل بینایی رو به رو هستند به صورت روزمره برای خواندن متون مختلف نیاز به کمک دارند؛ به عنوان مثال افراد سالمند که معمولاً برای خواندن متن درج شده روی داروهای خورد دچار مشکل هستند می‌توانند از سیستم‌های استخراج متن از تصویر استفاده کنند.

- **کمک به گردشگرها:** گردشگرانی که به کشورهای خارجی سفر می‌کنند معمولاً در پیدا کردن آدرس و خواندن متونی که به زبان کشور مقصد نوشته شده‌اند دچار مشکل می‌شوند؛ چرا که بیشتر اوقات با زبان کشور مقصد آشنا نیستند. به عنوان مثال گردشگری را تصویر کنید که به ایران سفر کرده و با زبان فارسی نیز آشنا نباشد؛ در این صورت بیشتر اوقات برای ترجمه متونی که به فارسی نوشته شده‌اند دچار مشکل خواهد شد. استفاده از روش‌هایی مانند سایت‌های ترجمه متن نیز به دلیل عدم آشنایی افراد با الفبای فارسی کارساز نخواهد بود.

- **شاخص‌گذاری و جستجو در منابع تصویری:** جستجو در منابع تصویری گاهی بسیار مشکل است. به عنوان مثال برخی اوقات برنامه‌های تلویزیونی نیاز به بررسی آرشیو تصویری بسیار بزرگی برای یافتن موردی خاص دارند. سیستم‌های استخراج متن از تصویر در این موارد نیز می‌توانند بسیار راهگشا باشند. به عنوان مثال فرض کنید در مورد شخصی در حال ساختن یک مستند هستیم و نیاز به برنامه‌هایی داریم که شخص مورد نظر در آن‌ها شرکت داشته است. با توجه به این که معمولاً اسم اشخاص پس از حضور آن‌ها در برنامه‌های تلویزیونی درج می‌شود، می‌توان از سیستم‌های استخراج



شکل ۱.۱: مراحل یک سیستم استخراج متن از تصویر

متنی برای یافتن ویدیوهای مورد نظر استفاده کرد.

- **کمک به خودروهای خودران:** خودروهای خودران برای رانندگی نیاز به اطلاعات متعددی دارند. یکی از این اطلاعات مربوط به علائم راهنمایی رانندگی است که در برخی موارد داخل آن‌ها متن درج شده است که سیستم‌های استخراج متن در این موارد می‌توانند بسیار راهگشا باشند.

۲.۱ تعریف مسئله و چالش‌های آن

سامانه‌های استخراج متن از تصاویر صحنه طبیعی را می‌توان به زیرسیستم‌های مکان‌یابی[۲]، استخراج[۳] و بهبود[۴]، و بازشناسی متن[۵] تقسیم کرد [۳]. هدف زیرسیستم مکان‌یابی متن، پیدا کردن و محدوده‌گذاری تمامی نواحی در یک تصویر (یا ویدئو) است که انسان آن‌ها را متن در نظر می‌گیرد [۴]. زیرسیستم استخراج و بهبود، متن را از پس‌زمینه جدا کرده و پس از بهبود کیفیت، آن را تبدیل به یک تصویر دودویی می‌کند و در انتها زیرسیستم بازشناسی متن، تصویر متن را به صورت دودویی دریافت کرده و متن موجود در آن را بازشناسی می‌کند. این مراحل در شکل ۱.۱ نشان داده شده است. مکان‌یابی متن را می‌توان مهم‌ترین مرحله در سیستم استخراج متن دانست؛ چرا که پیش‌نیاز مراحل بعدی است. پیدا کردن

[۲] Localization
[۳] Extraction
[۴] Enhancement
[۵] Text Recognition

متن در تصاویر صحنه طبیعی مسأله‌ای پرچالش و باز به شمار می‌رود. از جمله چالش‌های مکان‌یابی متون در تصاویر صحنه طبیعی می‌توان به موارد زیر اشاره کرد:

- **متغیر بودن اندازه، رنگ و فونت:** متون موجود در تصاویر به لحاظ اندازه، رنگ و فونت دارای تنوع زیادی هستند. این تنوع از آنجا ناشی می‌شود که این متون معمولاً توسط افراد مختلف، روی اشیای مختلفی و برای کاربردهای متفاوت مورد استفاده قرار می‌گیرند. این تنوع در ایران و زبان فارسی با توجه به سبک‌های مختلف نوشتاری بیشتر می‌شود. گاهی اوقات متون دارای فونت دقیقی نیستند و در پاره‌ای از مواقع نیز به صورت دستی نوشته می‌شوند.

- **پس‌زمینه ناهموار:** بر خلاف متون چاپی که در اسناد مورد شناسایی قرار می‌گیرند و دارای پس‌زمینه هموار هستند متون موجود در تصاویر صحنه طبیعی لزوماً دارای پس‌زمینه همواری نیستند و ممکن است رنگ‌هایی که در بخش‌های مختف متن در پس‌زمینه قرار دارند با هم متفاوت باشند که باعث دشوار شدن مکان‌یابی آن‌ها می‌شود.

- **چسبیده بودن نویسه‌های الفبای فارسی:** الفبای فارسی در برخی موارد به هم چسبیده هستند که باعث دشوارتر شدن مکان‌یابی متون فارسی از دو جهت می‌شود؛ از یک جهت باعث شباهت زیاد میان متون فارسی و شاخ و برگ درختان می‌شود و از جهت دیگر با توجه به اینکه بیشتر روش‌هایی که برای زبان‌هایی مانند انگلیسی ارائه شده‌اند روی یافتن کاراکترها در مرحله اول تکیه دارند به سادگی قابلیت اعمال روی زبان فارسی را ندارند.

۳.۱ اهداف و پیش‌فرض‌ها

تاکنون روش‌های زیادی برای مکان‌یابی متون غیرفارسی و به ویژه انگلیسی ارائه شده و موفقیت‌های زیادی نیز داشته‌اند؛ اما در زمینه مکان‌یابی متون فارسی پژوهش‌های زیادی انجام نگرفته است. روش‌های ارائه شده برای مکان‌یابی متون انگلیسی یا برای زبان فارسی

مناسب نیستند و یا نیاز به تغییرات گسترده‌ای دارند تا بتوانند برای زبان فارسی مورد استفاده قرار گیرند.

هدف از این پژوهش، مکان‌یابی متون فارسی در تصاویر صحنه طبیعی است؛ امّا روش‌های ارائه شده می‌توانند برای مکان‌یابی متون انگلیسی نیز مورد استفاده قرار گیرند. پیش‌فرض‌های روش‌های ارائه شده نیز به قرار زیر هستند:

- متون باید تقریباً به صورت افقی قرار داشته باشند.
- گاهی اوقات به دلیل شدت روشنایی بسیار کم، پس‌زمینه بسیار پیچیده، زاویه دید نامناسب و یا فاصله بسیار زیاد، انسان می‌تواند محل متن را حدس بزند امّا قادر به بازخوانی آن نیست. چنین متونی مد نظر ما نیستند.
- متون مورد نظر در این رساله شامل متونی که با اسپری و موارد مشابه نوشته می‌شوند و مرز مشخصی با پس‌زمینه ندارند یا دارای عرض قلم بسیار کم هستند نمی‌شود.
- در مورد فونت‌هایی که در زبان فارسی برای نوشتن متون مورد استفاده قرار می‌گیرند با توجه به نحوه کارکرد روش‌های ارائه شده، محدودیت خاصی در نظر گرفته نشده است؛ امَا گاهی اوقات در نوشتار فارسی به منظور زیبایی‌نویسی چندین کلمه و یا خط متنی به صورت درهم‌تنیده نوشته می‌شوند که مکان‌یابی این نوع از نوشتار نیز در این رساله مورد هدف نیست.

۴.۱ نوآوری‌های پژوهش

در این پژوهش چهار روش برای مکان‌یابی متون در تصاویر صحنه طبیعی ارائه می‌شود:

- روش اول با نام «امضای رنگ لبه‌ها» مفهومی به نام لبه‌های رنگی را معرفی کرده و با استفاده از آن به مکان‌یابی متون می‌پردازد. این روش پس از استخراج لبه‌های تصویر ورودی، با استفاده از الگوریتم انتقال میانگین[۶] رنگ‌های اطراف لبه‌ها را استخراج

[۶] Mean Shift

کرده و در ادامه با استفاده از الگوریتم خوشه‌بندی اتصال ـ تکی[۷] به گروه‌بندی نقاط لبه با استفاده از مکان قرار گرفتن آن‌ها می‌پردازند. این گروه‌ها در ادامه با استفاده از یک کلاسه‌بند آبشاری[۸] کلاسه‌بندی می‌شوند. همچنین تعدادی ویژگی جدید نیز برای استفاده در این کلاسه‌بند معرفی می‌شوند.

- روش دوم پیشنهادی با نام «تبدیل رنگ لبه‌ها» در چند بخش روش اول را بهبود می‌دهد. در این روش لبه‌های استخراج شده از تصویر بهبود داده می‌شوند. همچنین عملگری به نام «تبدیل رنگ لبه»[۹] برای به دست آوردن رنگ‌های اطراف لبه‌ها معرفی می‌شود که سریعتر و دقیق‌تر از روش پیشین است. همچنین برای گروه‌بندی نقاط لبه برای ایجاد نواحی کاندیدای متنی نیز در این فصل یک روش جدید معرفی می‌شود که بسیار سریع‌تر و بهتر از الگوریتم خوشه‌بندی اتصال ـ تکی عمل می‌کند.

- روش سوم با نام «تبدیل رنگ عمیق» به بهبود روش دوم پرداخته و برخی از مشکلات آن را رفع می‌نماید. همچنین در این روش یک شبکه عصبی کانولوشنی[۱۰] با ساختاری جدید برای جداسازی مناطق متنی و غیرمتنی معرفی می‌شود. در ادامه از این شبکه به صورت مستقل نیز برای مکان‌یابی متون استفاده می‌شود.

- خروجی شبکه کانولوشنی روش تبدیل رنگ عمیق عددی در بازه [۱...۰] است که درصد متن داخل ناحیه کاندیدا را نشان می‌دهد؛ چنین خروجی برخلاف روند معمول است که خروجی را به عنوان احتمال متن بودن ناحیه مورد نظرتعبیر می‌کنند. چنین تعبیر متفاوتی به ما اجازه می‌دهد که از این خروجی در روش چهارم با نام «شبکه فرا اکتشافی» استفاده کنیم. این روش یک الگوریتم فرااکتشافی جدید است که با طی چند گام ساده به مکان‌یابی متون در تصاویر می‌پردازد. این روش با پخش کردن چندین پنجره به صورت تصادفی در بخش‌های مختلف تصویر و استفاده از شبکه کانولوشنی به صورت تدریجی پنجره‌ها را به سوی بخش‌های متنی سوق می‌دهد.

---
[۷] Single Linkage
[۸] Cascade
[۹] Edge Color Transform
[۱۰] Convolutional Neural Network

۵.۱ ساختار رساله

ادامه رساله به صورت زیر سازمان‌دهی شده است. در فصل ۲ به بررسی کارهای انجام شده در این زمینه می‌پردازیم؛ فصل ۳ به معرفی مختصر یادگیری عمیق و شبکه‌های عصبی کانولوشنی می‌پردازیم؛ روش امضای رنگ لبه‌ها در فصل ۴ معرفی می‌شود؛ روش تبدیل رنگ لبه‌ها در فصل ۵ شرح داده می‌شود.جزئیات دو روش مبتنی بر یادگیری عمیق نیز در فصل ۶ توضیح داده شده است و در نهایت فصل ۷ به مقایسه روش‌های پیشنهادی با همدیگر و با چندین روش دیگر پرداخته و نتایج به دست آمده را جمع‌بندی می‌کند.

فصل ۲

مروری بر کارهای گذشته

۱.۲ مقدمه

متون موجود در تصاویر حاوی اطلاعات با ارزشی هستند که می‌توان از این اطلاعات برای تحلیل محتوای تصاویر استفاده کرد. با توجه به استفاده روزمره از دوربین گوشی‌های همراه و دیگر دستگاه‌های ضبط تصویر قابل حمل، تصاویر حاوی متون به سادگی به دست آمده و می‌توانند برای تحلیل محیط پیرامون مورد استفاده قرار گیرند؛ در نتیجه مکان‌یابی متون در تصاویر صحنه طبیعی تبدیل به یک موضوع تحقیقاتی فعال در بینایی ماشین شده است. اغلب روش‌هایی که قبل از سال ۲۰۱۵ میلادی در این زمینه ارائه شده‌اند مبتنی بر ویژگی‌های دست‌ساز هستند؛ امّا پس از این تاریخ به صورت تدریجی روش‌های مبتنی بر یادگیری عمیق تبدیل به جریان اصلی در این زمینه شدند. روش‌های موجود برای مکان‌یابی متن را می‌توان در دو دسته کلی جای داد [۴]:

- **روش‌های مبتنی بر پنجره لغزان:** روش‌های موجود در این گروه از یک پنجره استفاده

می‌کنند که روی تصویر حرکت داده می‌شود و حضور متن با استفاده از ویژگی‌های محلی استخراج شده از این پنجره تشخیص داده می‌شود. روش‌های مبتنی بر پنجره در حالت کلی مقاومت بیشتری نسبت به نویز دارند امّا معمولاً پیچیدگی محاسباتی آن‌ها به دلیل نیاز به جستجو با اندازه‌های مختلف پنجره، مقیاس‌های مختلف و احتمالاً چرخش‌های مختلف زیاد است [۵].

- **روش‌های مبتنی بر آنالیز مؤلفه‌های متصل**[۱]: روش‌های موجود در این دسته ابتدا به شناسایی کاراکترها می‌پردازند و سپس از کنار هم قرار دادن کاراکترها محدوده متن را مشخص می‌کنند. این روش‌ها معمولاً به لحاظ چگونگی شناسایی کاراکترهای جداگانه با هم تفاوت دارند که می‌تواند بر مبنای لبه‌ها، محاسبه مقدار انرژی کاراکتر و یا تشخیص نواحی حدی باشد. معمولاً گروه‌بندی این کاراکترها با استفاده از روش‌های اکتشافی و قوانین حاکم میان کاراکترهای متوالی انجام می‌شود. با توجه به اینکه کاراکترهای زبان فارسی به هم چسبیده هستند این روش‌ها در زبان انگلیسی کارایی بهتری نسبت به زبان فارسی خواهند داشت.

۲.۲ روش‌های مبتنی بر پنجره لغزان

پن و همکاران [۶] با استفاده از والدبوست [۷] و ویژگی‌های هیستوگرام گرادیان‌های جهت‌دار[۲] [۸] یک نقشه اطمینان متن روی هرمی از تصاویر خاکستری ایجاد می‌کنند. نواحی کاندید به صورت مستقل روی تصویر خاکستری با روش دودویی‌سازی[۳] نیبلاک [۹] تشخیص داده می‌شوند و زمینه تصادفی مشروط[۴] [۱۰] برای برچسب‌گذاری نواحی به عنوان متن یا غیر متن استفاده می‌شود که نقشه اطمینان متن به عنوان یکی از ورودی‌های این روش مورد استفاده قرار می‌گیرد. این روش به دلیل استفاده از زمینه تصادفی مشروط از لحاظ محاسباتی یک روش پرهزینه است.

---

[۱] Connected Component Analysis

[۲] Histogram of Oriented Gradients

[۳] Binarization

[۴] Conditional Random Field

چن و یولی [۱۱] از کلاسه‌بند آدابوست [۱۲] با ویژگی‌های میانگین شدت روشنایی، واریانس شدت روشنایی، ویژگی‌های هیستوگرامی استفاده کرده‌اند. پس از آن نسخه‌ای از الگوریتم دودویی‌سازی وفقی نیبلاک را برای به دست آوردن قطعه‌بندی مورد استفاده قرار داده‌اند. این روش به لحاظ محاسباتی پرهزینه بوده و نیاز به برش دستی تعداد زیادی زیر پنجره برای آموزش دارد. این روش توسط پن و همکاران [۱۳] با به کارگیری هیستوگرام گرادیان‌های جهت‌دار و الگوی دودویی محلی چند مقیاسه[۵] در مرحله تشخیص بهبود داده شده است. مینتو و همکاران [۱۴] گونه جدیدی از ویژگی هیستوگرام گرادیان‌های جهت‌دار را برای مکان‌یابی متن ارائه کرده‌اند که از مرزهای نرم (گوسی) به جای مرزهای سختی که در سلول‌های عادی این روش استفاده می‌شوند استفاده می‌کند.

ژائو و همکاران [۱۵] پس از استخراج ویژگی‌های توصیف‌گر $BRISK$[۶] [۱۶] از شبکه $SGONG$[۷] [۱۷] برای خوشه‌بندی توصیف‌گرها استفاده می‌کنند. در نهایت از مراکز خوشه برای تولید ویژگی‌های مربوط به روش کیسه کلمات بصری[۸] استفاده می‌شود. از طرف دیگر ویژگی‌های هیستوگرام کانال‌های رنگی نیز استخراج شده و این دو مجموعه ویژگی به عنوان ورودی یک کلاسه‌بندی آدابوست آبشاری برای تشخیص نواحی متنی مورد استفاده قرار می‌گیرند.

تیان و همکاران [۱۸] روش پنجره لغزان را با شبکه‌های عصبی کانولوشنی ترکیب کرده‌اند. در ابتدا نواحی کاندیدا با استفاده از یک پنجره لغزان که از یک کلاسه‌بند آبشاری استفاده می‌کنند مشخص می‌شوند؛ در ادامه یک شبکه عصبی کانولوشنی برای تعیین خطوط متنی استفاده می‌شود. تیان و همکاران [۱۹] برای تشخیص مناطق متنی یک پنجره لغزان به همراه یک شبکه «حافظه بلند کوتاه مدت دو جهته» [۹] [۲۰] را به خروجی شبکه کانولوشنی VGG [۲۱] اعمال می‌کنند. ژونگ و همکاران [۲۲] نیز روشی بر اساس شبکه کانولوشنی $Inception$ [۲۳] و پنجره لغزان ارائه کرده‌اند. در این روش پنجره‌های لغزان با مقیاس‌های مختلف روی خروجی شبکه کانولوشنی اعمال می‌شوند. همچنین ساختار پنجره‌ها به شکلی

---

[۵] Multi-Scale Local Binary Pattern
[۶] Binary Robust Invariant Scalable Keypoints
[۷] Self-Growing and Self-Organized Neural Gas
[۸] Bag of Visual Words
[۹] Bidirectional Long Short Term Memory

طراحی شده است است که با ساختار متن هماهنگی کاملی داشته و بتواند به بهترین نحو متون موجود در تصاویر را شناسایی کند. مهمترین عیب این روش آن است که قادر به تشخیص متون غیر افقی نیست.

۳.۲ روش‌های مبتنی بر آنالیز مؤلفه‌های متصل

یکی از اولین روش‌ها در این زمینه توسط اویا و همکارانش [۲۴] ارائه شد. آن‌ها از یک روش آستانه‌گیری وفقی محلی در تصاویر خاکستری برای تشخیص نواحی کاندید متن استفاده کردند و نواحی با کنتراست مناسب را به عنوان کاراکتر در نظر گرفتند. لی و همکاران [۲۵] ابتدا تصویر را به فضای تقطیع‌شده برده و در این فضا آستانه‌گذاری انجام می‌دهند و پس از آن کاراکترها را با قوانین ساده هم‌ترازی گروه‌بندی می‌کنند. در هر دو روش فرض شده است که کنتراست ناحیه متن نسبت به پس زمینه بالاست و پس‌زمینه یکنواخت است. کیم و همکاران [۲۶] سه روش مستقل تشخیص پیوستگی رنگ، تشخیص لبه و واریانس رنگ را برای یافتن نواحی نامزد با هم ترکیب کرده‌اند. تاکاهاشی و همکاران [۲۷] از یک روش مبتنی بر گراف استفاده کرده‌اند که در آن گره‌ها نشان دهنده کاراکترها و یال‌ها نشان‌دهنده رابطه میان کاراکترها در بلوک متن هستند. این روش نمی‌تواند تغییرات شدت روشنایی در متن و یا پس‌زمینه را مدیریت کند. همچنین یافتن مقدار وزن مناسب برای یال‌ها یک مسئله سخت به شمار می‌رود.

اپشتین و همکاران [۲۸] عملگر تبدیل عرض قلم[۱۰] را برای اولین بار معرفی کرده‌اند. در این روش عرض قلم برای هر پیکسل تصویر پس از به دست آوردن لبه‌ها تخمین زده می شود. پس از این کار الگوریتمی ساده برای یافتن مؤلفه‌های متصل و ایجاد نواحی نامزد کاراکتر استفاده می‌شود؛ این روش در مقاله‌های [۲۹] و [۳۰] بهبود داده شده است. روش ارائه شده در [۳۱] نیز از عرض قلم استفاده می‌کند. این روش با الهام از الگوریتم تشخیص گوشه[۱۱] *Fast* [۳۲] دو عملگر برای تشخیص نقاط انتهایی و نقاط انحنای کاراکترها معرفی

---

[۱۰] Stroke Width Transform

[۱۱] Corner Detection

کرده است. پس از تشخیص این نقاط، با استفاده از یک روش آستانه‌گیری محلی، نواحی کاندیدای کاراکتری استخراج می‌شوند؛ در ادامه از این نواحی ویژگی‌های هندسی و عرض قلم استخراج شده و به یک کلاسه‌بند آدابوست *Gentle* [۳۳] برای جدا کردن کاراکترهای متنی داده می‌شود. در انتها، کاراکترهای تشخیص داده شده با استفاده از یک الگوریتم خوشه‌بندی کنار هم قرار گرفته و نواحی متنی را ایجاد می‌کنند.

خوشه‌بندی یکی دیگر از روش‌هایی است که برای یافتن نواحی احتمالی متن مورد استفاده قرار گرفته است. قوشال و همکاران [۳۴]برای خوشه‌بندی تصویر و به دست آوردن نواحی کاندید متن از روشی مبتنی بر خوشه‌بندی فازی و ویژگی‌های واریانس و قدرت لبه‌ها در پنجره‌های محلی استفاده کرده‌اند. کوتس و همکاران [۳۵] از یک روش یادگیری ماشین بدون ناظر برای تشخیص و شناسایی کاراکترها استفاده کرده‌اند. یک پنجره ۳۲×۳۲ پیکسل با چندین مقیاس روی تصویر لغزانده می‌شود و هر پنجره با استفاده از یک کلاسه‌بند ماشین بردار پشتیبان با هسته خطی به متن و غیر متن برچسب‌گذاری می‌شود. یان و همکاران [۳۶] نیز روشی بر پایه الگوریتم خوشه‌بندی فازی ارائه کرده‌اند. روش دیگری که در [۳۷] پیشنهاد شده است استفاده از خوشه‌بندی برای تقسیم‌بندی تصویر به تعدادی خوشه و پس از آن گروه‌بندی خوشه‌های یافت شده با قوانین اکتشافی است. این روش فرض می‌کند که متن و پس‌زمینه دارای رنگ یکنواخت هستند. از دیگر روش‌های مبتنی بر خوشه‌بندی می‌توان به روش ارائه شده در مرجع [۳۸] اشاره کرد.

یائو و همکاران [۳۹] با استفاده از روش اختصاصی خود ابتدا تصویر ورودی را با آستانه‌گذاری وفقی به یک تصویر خاکستری با ۳ سطح روشنایی ۰، ۱۰۰ و ۲۵۵ تبدیل کرده‌اند. پس از آن نقاطی که دارای سطح خاکستری ۰ یا ۲۵۵ هستند به عنوان نقاط کاندید متن در نظر گرفته می‌شوند. پس از آن با استفاده از یک سری خواص ساختاری مؤلفه‌های متصل، به آن‌ها برچسب متن و یا غیر متن داده‌اند. فابریزیو و همکاران [۴۰] یک عملگر مورفولوژی جدید برای یافتن نواحی متنی استفاده کرده‌اند. لو و همکاران [۴۱] لبه‌های تصویر ورودی را به دست آورده و با استفاده از ویژگی‌هایی مانند کنتراست، عرض لبه و شمارنده برش‌های عمودی و افقی لبه‌ها، تعدادی از بخش‌های غیر متنی را حذف می‌کنند. پس از آن با استفاده

از آستانه‌گیری محلی مناطق کاندیدای کاراکتر را به دست می‌آورند؛ در نهایت با استفاده از کیسه کلمات بصری[۱۲] و ماشین بردار پشتبان مناطق کاندیدای کاراکتری کلاسه‌بندی می‌شوند. جدنبرگ و همکاران [۴۲] از روش‌های جعبه‌های لبه[۱۳] [۴۳] و ویژگی‌های کانال تجمیع شده[۱۴] [۴۴] برای تولید نواحی کاندیدا استفاده کرده و نتیجه نهایی را با استفاده از جنگل تصادفی [۴۵] و شبکه عصبی کانولوشنی هرس کرده‌اند. هی و همکاران [۴۶] نیز در ابتدا لبه‌های تصویر ورودی را به دست آورده و سعی می‌کنند تا لبه‌های اشیای مختلف را از هم تفکیک کنند و در ادامه با استفاده از یک شبکه کانولوشنی نواحی متنی را از غیر متنی جدا می‌کنند. هوانگ و همکاران [۴۷] از ترکیب «نقشه برجستگی»[۱۵] و لبه‌ها برای ایجاد یک روش مقاوم برای مکان‌یابی متون در تصاویر با کیفیت پایین استفاده کرده‌اند. درنهایت نواحی به دست آمده از مرحله قبل با استفاده از یک «مدل ترکیب گوسی»[۱۶] [۴۸] به کاراکترهای سفید، سیاه و پس‌زمینه تقسیم می‌شوند. در ادامه با استفاده از تعدادی ویژگی هندسی این نواحی گروه‌بندی شده و بخش‌های متنی به دست می‌آیند. دنگ و همکاران [۴۹] از ساختاری مشابه به شبکه SSD[۱۷] [۵۰] برای استخراج ویژگی و تعیین پیکسل‌های متنی و نحوه اتصال آن‌ها استفاده کرده‌اند؛ آن‌ها در ادامه با اتصال پیکسل‌های متنی کاراکترها را ساخته و با اتصال کاراکترها خطوط متنی را تعیین کرده‌اند.

شی و همکاران [۵۱] روشی را ارائه کرده‌اند که مبنای آن استفاده از «نواحی حدی پایدار بیشینه»[۱۸] است. یک «ناحیه حدی» یک مؤلفه متصل است که پیکسل‌های آن شدت روشنایی بیشتر و یا کمتر از محدوده بیرونی‌اش داشته باشد [۵۲] و [۵۳]. در این روش ابتدا نواحی حدی پایدار از تصویر استخراج می‌شود. پس از آن گرافی تشکیل می‌شود که رئوس آن را نواحی حدی تشکیل می‌دهند. در ادامه جریمه‌ای برای اعمال برچسب متن و یا غیر متن به هر یک از رئوس این گراف در نظر گرفته می‌شود و الگوریتم سعی می‌کند

[۱۲] Bag Of Visual Words
[۱۳] Edge Boxes
[۱۴] Aggregated Channel Features
[۱۵] Saliency Map
[۱۶] Gaussian Mixture Model
[۱۷] Single-Shot Detector
[۱۸] Maximally StableExteremal Region

تا میزان این جریمه را برای کل گراف کمینه کند. این روش با توجه به قوانین اکتشافی که در آن استفاده شده است محدودیت‌های زیادی دارد. لی و همکاران [۵۴] نیز روشی را ارائه کرده‌اند که بسیار شبیه رویکرد شی و همکاران [۵۱] است. تنها تفاوت این دو روش در ویژگی‌هایی است که به کار برده شده است، مثلاً به جای محاسبه ویژگی‌های هیستوگرام گرادیان جهت‌دار روی تمام یک ناحیه، این ویژگی‌ها فقط روی لبه‌هایی محاسبه شده‌اند که توسط الگوریتم استخراج لبه کنی استخراج شده است. یین و همکاران [۵۵] نیز روشی بر مبنای نواحی حدی پایدار بیشینه ارائه کرده‌اند. در این روش پس از استخراج نواحی نامزد متن، مقداری به عنوان جریمه به ناپایداری نواحی که مقدار ناپایداری آن‌ها از محدوده خاصی تجاوز می‌کند اضافه می‌شود. در ادامه در درخت نواحی حدی پایدار بیشینه، بین پدر و فرزندش تنها ناحیه‌ای انتخاب می‌شود که مقدار ناپایداری کمتری داشته باشد. پس از حذف نواحی اضافی، با استفاده از شباهت ویژگی‌های ساختاری نواحی مانند رنگ و نسبت طول و عرض مستطیل محیط و استفاده از الگوریتم خوشه‌بندی [۱۹] [۵۶] نواحی تشخیص داده شده کنار هم قرار گرفته و خطوط متن را تشکیل می‌دهند. این روش در [۵۷] برای مکان‌یابی نواحی متنی غیر افقی بهبود داده شده است. دائی و همکاران [۵۸] پس از بهبود کیفیت تصویر، نواحی حدی پایدار بیشینه را برای کانال‌های مختلف رنگی محاسبه کرده‌اند. روش ارائه شده در [۵۹] هم از دیگر روش‌هایی است که از نواحی حدی پایدار بیشینه به همراه تعدادی ویژگی هندسی برای حذف نواحی غیرمتنی استفاده کرده است. ونگ و همکاران [۶۰] نواحی حدی پایدار بیشینه را با استفاده از شبکه عصبی کانولوشنی به متن و غیر متن تفکیک کرده‌اند.

ژنگ و همکاران [۶۱] پس از استخراج نواحی حدی پایدار بیشینه، نواحی را که دارای رنگ یکنواختی نیستند حذف می‌کنند. در ادامه پس از استخراج لبه‌های تصویر و به دست آوردن لبه‌های مخصوص به هر ناحیه حدی پایدار بیشینه، عرض قلم برای هر یک از نواحی محاسبه می‌شود. در ادامه از یک شبکه عصبی کانولوشنی برای کلاسه‌بندی نواحی به کاراکتر و غیر کاراکتر استفاده می‌شود. در نهایت با استفاده از یک الگوریتم خوشه‌بندی کاراکترها در کنار

---

[۱۹] Single-Linkage

هم قرار گرفته و ناحیه متنی را ایجاد می‌کنند.

سونی و همکاران [۶۲] با استفاده از «فیلتر هدایت‌شده سریع»[۲۰] [۶۳] نواحی حدی پایدار بیشینه را بهبود داده‌اند تا کارایی بهتری در تصاویر محو[۲۱] داشته باشد و از اتصال کاراکترهایی که به همدیگر نزدیک هستند جلوگیری شود. در ادامه ترکیبی از ویژگی‌های انحراف عرض قلم[۲۲]، برای اندازه‌گیری یکنواختی عرض قلم در مؤلفه‌های متصل، هیستوگرام گرادیان جهت‌دار، ویژگی‌های مربوط به رنگ و چندین ویژگی هندسی به یک کلاسه‌بند «بیز ساده»[۲۳] [۶۴] برای تشخیص مؤلفه‌های متنی داده می‌شوند؛ در نهایت الگوریتم خوشه‌بندی انتقال میانگین برای گروه‌بندی نواحی متنی به کار گرفته می‌شود. «نواحی حدی»[۲۴] روش مشابه دیگری است که در چندین مقاله از جمله [۶۵]، [۶۶] و [۶۷] برای به دست آوردن نواحی کاندیدای کاراکتر مورد استفاده قرار گرفته است. «نواحی حدی» نسبت به کنتراست کم، تغییرات روشنایی و تغییرات رنگی مقاوم‌تر از نواحی حدی پایدار بیشینه هستند. نئومن و همکاران [۶۶] پس از استخراج نواحی حدی در ابتدا با استفاده از یک کلاسه‌بند آدابوست، نواحی حدی غیر کاراکتری را حذف می‌کنند. در ادامه با استفاده از ویژگی‌های هندسی و مدل‌های زبانی که احتمال توالی کاراکترها را در زبان انگلیسی در نظر می‌گیرند به ساختن نواحی متنی می‌پردازند.

ژنگ و همکاران [۶۷] پس از استخراج نواحی حدی از چندین کانال رنگی مختلف، ترکیبی از ویژگی‌های هندسی، هیستوگرام گرادیان جهت‌دار و الگوی دودویی محلی را برای هرس کردن نواحی حدی غیر کاراکتری به کار برده‌اند؛ در ادامه با استفاده از اطلاعات مربوط به فاصله بین کاراکترها نواحی متنی تشکیل شده و از یک شبکه عصبی کانولوشنی برای هرس کردن الگوهای تکرار شونده مانند نرده‌ها و آجر استفاده می‌کنند.

شبکه‌های کاملاً کانولوشنی [۲۵] [۶۸] نیز در سال‌های اخیر مورد توجه قرا گرفته‌اند [۶۹–۷۱].

---

[۲۰] Fast Guided Filter

[۲۱] Blurred

[۲۲] Stroke Width Deviation

[۲۳] Naive Bayes

[۲۴] Extremal Regions

[۲۵] Fully Convlutional Networks

روش‌های مبتنی بر شبکه‌های کاملاً کانولوشنی، لایه‌های کاملاً‌ـ‌متصل انتهایی شبکه‌های کانولوشنی را با لایه‌های کانولوشنی جایگزین می‌کنند؛ این کار به این روش‌ها اجازه می‌دهد تا اطلاعات کلی مربوط به کاراکترها را در لایه‌های آخر حفظ کند که برای مکان‌یابی متون مهم است. ژنگ و همکاران [۷۰] روش برچسب‌گذاری معنایی توسط شبکه‌های کانولوشنی را با خروجی نواحی حدی پایدار بیشینه ترکیب کرده‌اند تا بتوانند روشی برای مکان‌یابی متون با جهت دلخواه ایجاد کنند. هی و همکاران [۷۲] نیز روش مشابهی بر پایه رگرسیون در این زمینه ارائه کرده‌اند. ژنگ و همکاران [۷۱] نیز روش دیگری بر اساس شبکه‌های کاملاً کانولوشنی ارائه کرده‌اند که به صورت همزمان هم عددی به عنوان امتیاز کاراکتر بودن تولید می‌کند و هم محل پیش‌بینی شده برای کاراکترها را بهبود می‌دهد. از این امتیاز تولید شده در مراحل بعدی برای رتبه‌بندی مناطق کاندیدای متنی و حذف مناطق غیرمتنی استفاده می‌شود. در روش دیگری گوپتا و همکاران [۶۹] با ترکیب شبکه‌های کاملاً کانولوشنی و شبکه مکان‌یابی اشیای ارائه شده در [۷۳] روش جدیدی ارائه کرده‌اند که به صورت همزمان هم مکان متن و هم احتمال متن بودن مناطق مختلف تصویر را پیش‌بینی می‌کند. بزازیان و همکاران [۷۴] شبکه‌های کاملاً کانولوشنی را با شبکه تولید ناحیه کاندیدای ارائه شده توسط گومز و همکاران [۷۵] ترکیب کرده‌اند. در این روش در ابتدا یک نقشه اطمینان متنی توسط شبکه کاملاً کانولوشنی ایجاد می‌شود؛ این اطلاعات در مرحله بعدی توسط شبکه تولید ناحیه کاندیدا مورد استفاده قرار گرفته و نواحی کاندیدا تولید می‌شود. نواحی کاندیدای تولید شده در نهایت با کمک نواحی حدی پایدار بیشینه هرس می‌شوند تا مناطق نهایی متن به دست بیاید.

یکی از اولین روش‌های ارائه شده در زبان فارسی نیز، روش ارائه شده توسط داراب و همکاران [۷۶] است. در این روش ابتدا با استفاده از روش تشخیص لبه سوبل لبه‌ها استخراج می‌شوند و پس از آن با استفاده از عملگر مورفولوژی دایلیشن[۲۶] لبه‌ها ضخیم‌تر می‌شوند. تشخیص لبه‌ها نیز به صورت وفقی انجام می‌شود. پس از اعمال عملگرهای مورفولوژی، مناطقی که در آن‌ها لبه‌های عمودی و افقی به هم متصل شده‌اند به عنوان نقاط کاندید برای متن در

[۲۶] Dilation

نظر گرفته می‌شوند. پس از آن خطوط کاندید برای متن استخراج می‌شوند. در ادامه از یک سری قوانین اکتشافی مانند نسبت طول به عرض ناحیه برای هرس کردن بیشتر نواحی استفاده می‌شود. این روش برای تشخیص متن در اندازه‌های مختلف، هرمی متشکل از ۷ لایه از تصویر را ساخته و مراحل قبلی را روی هر ۷ تصویر اعمال می‌کند و در حالاتی که متن در بیش از یک سطح تشخیص داده شود از کلاسه‌بند ماشین بردار پشتیبان برای رفع مشکل استفاده می‌شود. در مواردی که متن داخل تصویر دارای فونت‌های بسیار کوچک یا بسیار بزرگ باشد، این روش با مشکل مواجه می‌شود. همچنین اگر در تصویر شاخ و برگ درختان وجود داشته باشد این روش در تمایز بین متن فارسی و شاخ و برگ دچار مشکل می‌شود.

قانعی و همکاران [۷۷] نیز عملگری با نام «نواحی با عرض قلم پایدار»[۲۷] برای تعیین نواحی کاندیدای متن ارائه کرده‌اند. این عملگر بر اساس این فرض کار می‌کند که نواحی متنی دارای مرزهای بسته بوده و عرض قلم آن‌ها یکنواخت است. این عملگر به عنوان فیلتری برای هرس کردن خروجی نواحی حدی پایدار بیشینه مورد استفاده قرار گرفته است. این روش نمی‌تواند نواحی متنی را که در مرزهای تصویر ورودی قرار گرفته‌اند مکان‌یابی کند. همچنین بسیاری از نواحی متنی تکرار شونده مانند آجر و پنجره‌ها را به عنوان متن شناسایی می‌شوند. این روش در [۷۸] بهبوده داده شده است. در این روش نیز پس از استخراج نواحی حدی پایدار بیشینه تعدادی از آن‌ها با در نظر گرفتن یکنواختی عرض قلم در مناطق متنی هرس می‌شوند. در ادامه ویژگی‌های رنگ، کنتراست، لبه و تعدادی ویژگی هندسی برای ایجاد یک سیستم فازی برای هرس کردن نواحی غیرمتنی مورد استفاده قرار می‌گیرد. همچنین برای هرس کردن نواحی غیر متنی در صحنه‌های پیچیده‌تر مدل «بازیابی ماتریس رده- پایین»[۲۸] [۷۹] که روشی برای ایجاد نقشه برجستگی است مورد استفاده قرار گرفته است. در این مدل تصویر ورودی با یک ماتریس رده- پایین به اضافه نویزهای خلوت[۲۹] در فضای ویژگی تخمین زده می‌شود که در این مورد ماتریس رده- پایین نشان‌دهنده پس‌زمینه

[۲۷] Stable Width Stroke Regions
[۲۸] Low-Rank Matrix Recovery
[۲۹] Sparse

تصویر و نویزهای خلوت نشان‌دهنده متن موجود در تصویر هستند. در نهایت ویژگی‌های مربوط به رنگ، عرض قلم و تعدادی ویژگی هندسی برای ایجاد نواحی متنی از کاراکترها مورد استفاده قرار می‌گیرند.

فصل ۳

یادگیری عمیق

۱.۳ مقدمه

یادگیری عمیق شاخه خاصی از یادگیری ماشین است که در سال‌های اخیر به شدت مورد توجه جامعه یادگیری ماشین قرار گرفته و در کاربردهای مختلفی به کار گرفته شده است [۸۰–۸۴]. در این رویکرد، یادگیری در لایه‌های متوالی به صورت سلسله مراتبی انجام می‌گیرد؛ در هر لایه بازنمایی‌های جدیدی از تصویر ورودی ساخته می‌شوند که به لحاظ ارتباط با خروجی مورد نظر، پر بارتر از لایه‌های پیش از خود هستند. کلمه عمیق در یادگیری عمیق هیچ ارتباطی با درک عمیق‌تری که از طریق آن حاصل می‌شود ندارد و لفظ عمیق در اینجا به این واقعیت اشاره دارد که معمولاً در این مدل‌ها تعداد زیادی لایه‌های متوالی در امتداد یکدیگر قرار گرفته و یک مدل را تولید می‌کنند [۱].

بیش از دو دهه از معرفی تکنیک‌های یادگیری عمیق و استفاده عملی از آن‌ها می‌گذرد [۸۵]؛ امّا دلیل رونق گرفتن استفاده از روش‌های یادگیری عمیق، افزایش قابل توجه

سرعت پردازش رایانه‌ها و نیز افزایش داده‌های در دسترس در سال‌های اخیر بوده است.
یکی از معروفترین مدل‌های مورد استفاده در یادگیری عمیق شبکه‌های عصبی کانولوشنی[۱] هستند. این شبکه‌ها از نحوه کارکرد قشر بینایی[۲] انسان الهام گرفته شده‌اند که در آن نرون‌های مشخص به ساختارهای متفاوتی حساس هستند و به آن‌ها واکنش نشان می‌دهند. تجمیع و پردازش این واکنش‌ها در مغز انسان باعث ایجاد توانایی بینایی در انسان می‌شود.

۲.۳ مزایای استفاده از شبکه‌های کانولوشنی

یکی از مهمترین دلایل محبوبیت شبکه‌های عصبی کانولوشنی، کارایی بالای آن‌ها در بسیاری از مسائل دنیای واقعی است که تا پیش از رونق گرفتن آن‌ها توسط روش‌های سطحی[۳] مانند ماشین بردار پشتیبان حل می‌شدند. امّا در سال‌های اخیر این شبکه‌ها در مسائل مختلفی مانند دسته‌بندی تصاویر، تشخیص گفتار، ترجمه ماشین و خودروهای خودران مورد استفاده قرار گرفته و به نتایج بسیار خوبی دست یافته‌اند [۸۶–۸۸].
مزیت دیگر شبکه‌های کانولوشنی عدم نیاز آن‌ها به مهندسی ویژگی است؛ روش‌های سطحی معمولاً نیاز به استخراج ویژگی از تصویر ورودی و همچنین پیش‌پردازش آن‌ها دارند؛ در حالی که یادگیری عمیق تا حد زیادی این نیاز را مرتفع می‌سازد. نکته دیگری که در این زمینه مهم است روند یادگیری ویژگی‌ها در یادگیری عمیق است که بازنمایی‌های[۴] متوالی داده‌ها را به صورت همزمان و در ارتباط با مسئله مورد نظر یاد می‌گیرد؛ این درست در نقطه مقابل روش‌های یادگیری سطحی قرار دارد که ویژگی‌ها معمولاً به صورت جداگانه استخراج می‌شوند و در ادامه یک کلاسه‌بند با استفاده از ویژگی‌های استخراج شده آموزش داده می‌شود. این روش یادگیری همزمان در شبکه‌های کانولوشنی به مدل آموزش داده شده این امکان را می‌دهد که بتواند ویژگی‌های بهتری را یاد بگیرد که کاملاً با مسئله مورد نظر مرتبط هستند.

---

[۱] Convolutional Neural Networks
[۲] Visual Cortex
[۳] Shallow
[۴] Reresentation

همانگونه که اشاره شد شبکه‌های عصبی کانولوشنی معمولاً از تعداد زیادی لایه متوالی که در امتداد همدیگر قرار گرفته‌اند برای یادگیری استفاده می‌کند. برای یادگیری پارامترهای لایه‌های این شبکه‌ها نیز از گونه‌های مختلف الگوریتم گرادیان نزولی [۵] استفاده می‌شود. در اینجا تعدادی از این لایه‌ها و نحوه کارکرد آن‌ها را به صورت مختصر مورد بررسی قرار می‌دهیم.

۳.۳ لایه کانولوشن

برخلاف شبکه‌های عصبی سطحی همانند شبکه پرسپترون چند لایه[۶] که حتی ورودی تصویر را به صورت برداری دریافت می‌کنند، ورودی شبکه‌های کانولوشنی به صورت چند یا یک کانال رنگی است. این نحوه کارکرد باعث می‌شود تا رابطه میان پیکسل‌های مجاور در اثر تبدیل تصویر به بردار از بین نرود. لایه کانولوشنی سعی می‌کند تا با استفاده از تعدادی فیلتر کوچک، که به آن‌ها کرنل[۷] نیز گفته می‌شود، الگوهای کلی را در مورد ساختارهای موجود در تصاویر ورودی بیاموزد.

استفاده از این فیلترهای کوچک دو ویژگی بسیار مهم را برای شبکه‌های کانولوشنی به ارمغان می‌آورد:

- الگوهایی که توسط این فیلترها یاد گرفته می‌شوند نسبت به جابه‌جایی حساس نیستند؛ به عنوان مثال اگر الگویی به خطوط مورب در یک جهت خاص حساس باشد در همه جای تصویر این حساسیت را نشان داده و خروجی مورد نظر را تولید خواهد کرد و محل قرار گرفتن خط مورب در داخل تصویر اهمیتی نخواهد داشت. چنین ساختاری به شبکه‌های کانولوشنی اجازه می‌دهد تا بتوانند با مشاهده نمونه‌هایی از یک کلاس به یادگیری ساختار آن کلاس بپردازند و در ادامه قادر باشند تا الگوهای مشاهده شده در این نمونه‌ها را با تصاویر جدید تطبیق دهند.

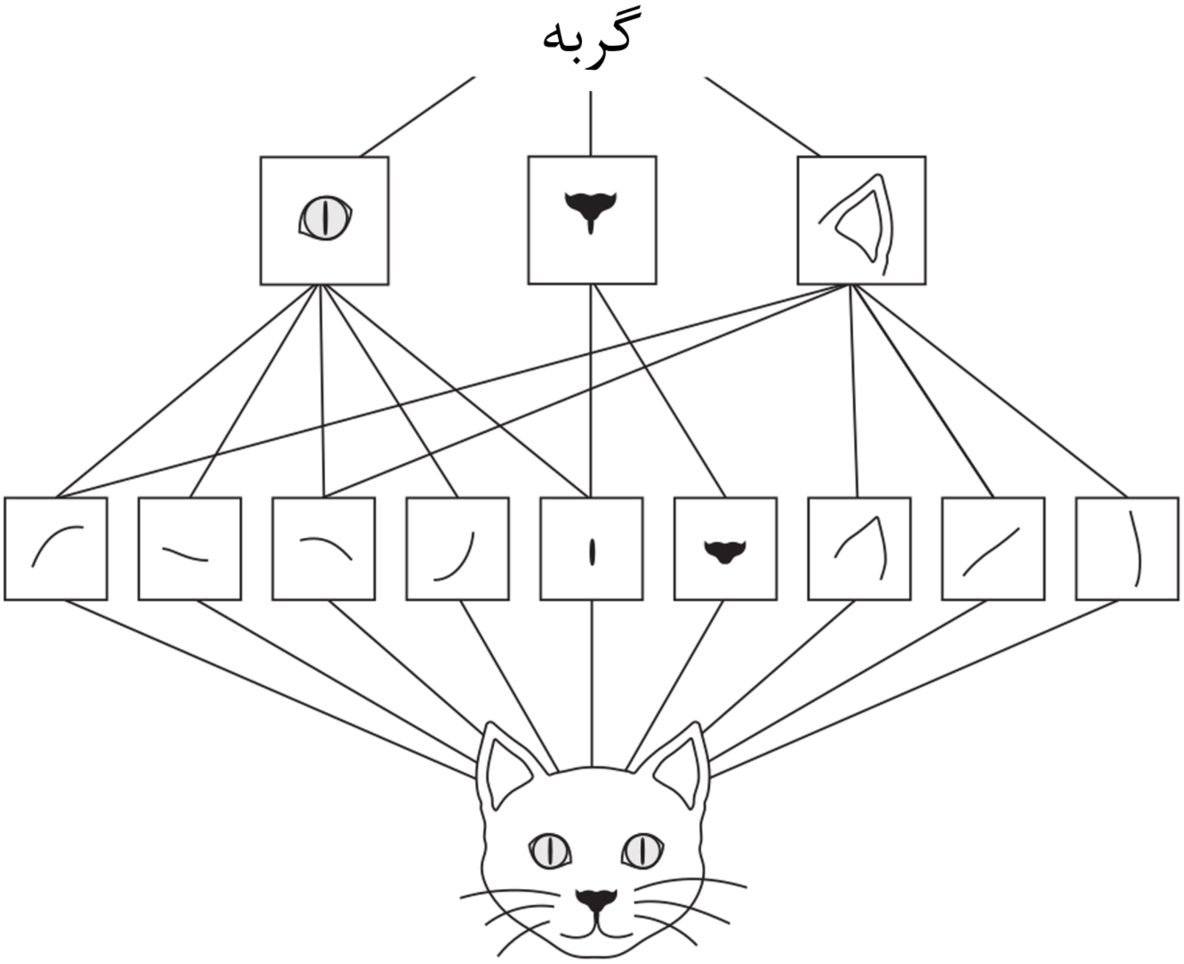
---

[۵] Gradient Descent

[۶] Multi-Layer Perceptron

[۷] Kernel

- این ساختار به شبکه‌های کانولوشنی اجازه می‌دهد تا الگوهای سلسله مراتبی را بیاموزند. به عنوان مثال ممکن است لایه‌های اولیه کانولوشنی به الگوهای کوچک محلی مانند لبه‌ها حساس باشند و لایه بعدی بتواند از ویژگی‌های لایه قبلی استفاده کرده و ویژگی‌های پیچیده‌تری را بیاموزد. این ساختار به این شبکه‌ها اجازه یادگیری مدل‌هایی را می‌دهد که به تدریج از تصویر ورودی یک ساختار پیچیده و انتزاعی را ایجاد می‌کنند. این موضوع در شکل ۱.۳ نشان داده شده است.

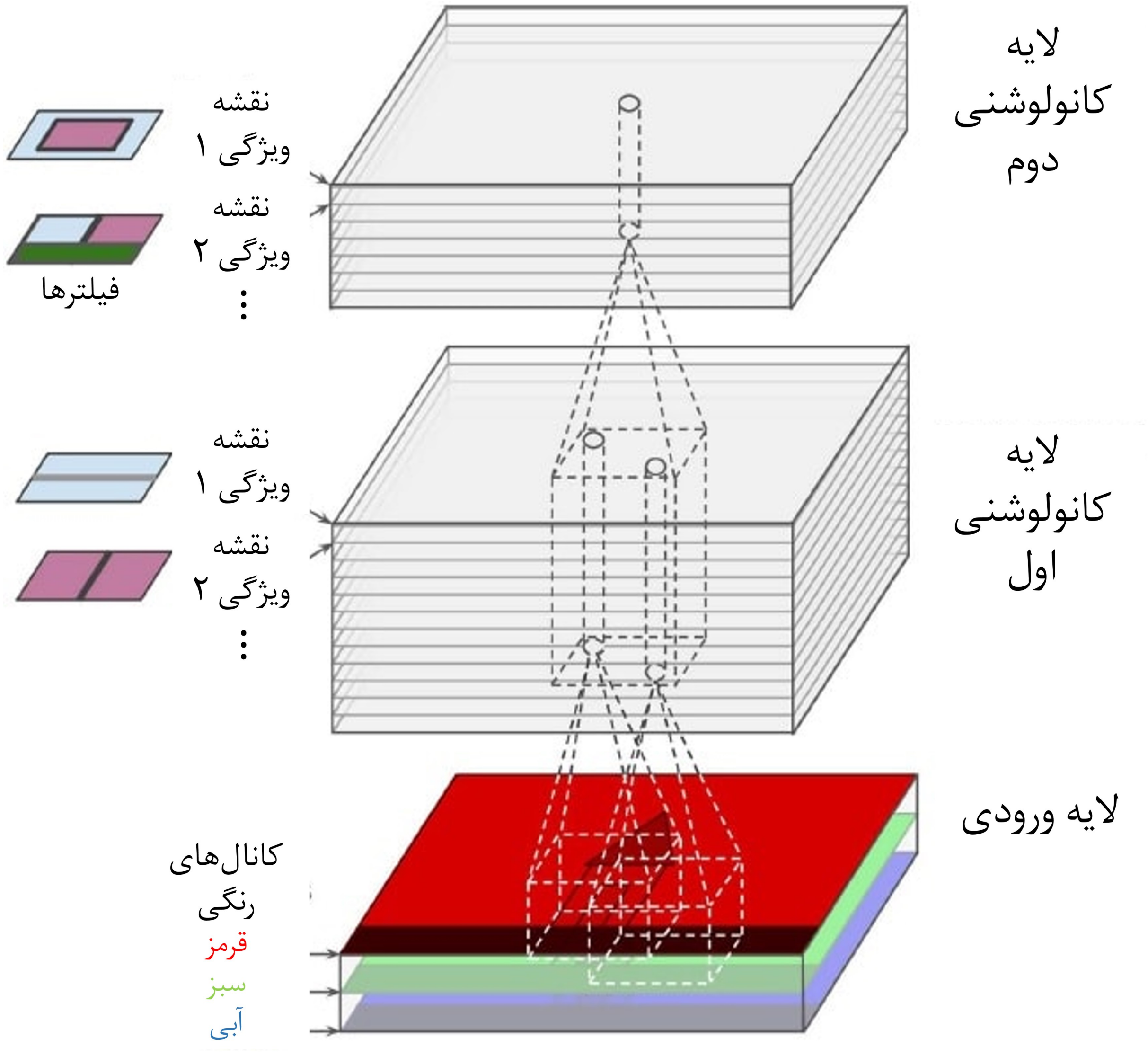


شکل ۱.۳: یادگیری ساختار سلسله مراتبی توسط شبکه‌های کانولوشنی [۱]

یک لایه کانولوشن شامل مجموعه‌ای از فیلترهای مستقل است؛ هر یک از این فیلترها به تعداد کانال‌های ورودی لایه، عمق دارند و طول و عرض آن‌ها در یک لایه با هم یکسان است. این فیلترها هر یک به صورت مستقل با نقشه ویژگی ورودی کانوالو[۸] می‌شوند و یک

[۸] Convolve

نقشه ویژگی به عنوان خروجی تولید می‌کنند؛ در ادامه این نقشه‌های ویژگی روی هم قرار گرفته و نقشه ویژگی ورودی لایه بعدی را می‌سازند. معمول‌ترین اندازه برای این فیلترها، پنجره‌های ۳ × ۳ هستند. جزئیات این موضوع نیز در شکل ۲.۳ آمده است.



شکل ۲.۳: نحوه ساخت نقشه‌های ویژگی در شبکه‌های کانولوشنی با استفاده از فیلترها که به صورت خطچین نشان داده شده‌اند.

۴.۳ لایه‌ی ادغام بیشینه

معمولاً در میان لایه‌های متوالی کانولوشن از لایه‌های ادغام بیشینه[۹] استفاده می‌شود [۸۹]؛ لایه ادغام بیشینه به صورت یک پنجره لغزان به هر یک از کانال‌های نقشه ویژگی ورودی اعمال می‌شود و هر بار بیشترین مقدار موجود در پنجره فیلتر را به عنوان خروجی باز می‌گرداند؛ معمولاً این پنجره‌ها با اندازه ۲ × ۲ و گام حرکت ۲ به کار برده می‌شوند. با توجه به اینکه این لایه به صورت مستقل روی هر یک از کانال‌های ورودی اعمال می‌شود، بنابراین تعداد کانال‌های ورودی با اعمال این لایه تغییر پیدا نمی‌کند امّا طول و عرض آن کاهش پیدا می‌کند.

دلایل استفاده از لایه ادغام بیشینه را می‌توان در دو مورد خلاصه کرد:

- کاهش متوالی ابعاد کانال‌های ورودی لایه‌های کانولوشنی باعث می‌شود که به تدریج ویژگی‌ها و الگوهای یادگرفته شده در لایه‌های قبلی در کنار یکدیگر قرار بگیرند؛ این همجواری به شبکه کمک می‌کند که در هر لایه ویژگی‌های سطح بالاتری نسبت به لایه‌های پیشین بیاموزد و در نهایت همانگونه که قبلاً نیز اشاره شد، یک مدل سطح بالا نسبت به تصویر ورودی ایجاد نماید.

- دلیل دیگر استفاده از لایه‌های ادغام، کاهش تعداد پارامترهای شبکه است؛ مثلاً در صورتی که از پنجره‌های با اندازه ۲ × ۲ با گام‌های ۲ استفاده کنیم، ابعاد هر یک از کانال‌های ویژگی ورودی، ۷۵ درصد کاهش پیدا می‌کند و این به نوبه خود پارامترهای شبکه را کاهش می‌دهد. این کاهش کمک شایانی به جلوگیری از بیش‌برازش[۱۰] شبکه و یادگیری بهتر آن می‌نماید.

در کنار لایه ادغام بیشینه، لایه‌های دیگری مانند ادغام میانگین[۱۱] نیز وجود دارند که به دلیل عملکرد ضعیف‌ترشان کمتر مورد استفاده قرار می‌گیرند؛ نحوه عملکرد لایه ادغام بیشینه

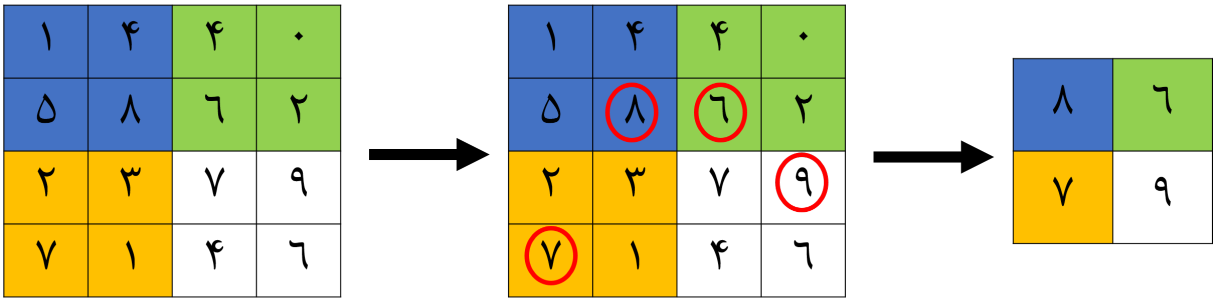
---

[۹] Max-Pooling Layer

[۱۰] Overfitting

[۱۱] Average Pooling

در شکل ۳.۳ با یک مثال نشان داده شده است.



شکل ۳.۳: نحوه کارکرد لایه ادغام بیشینه

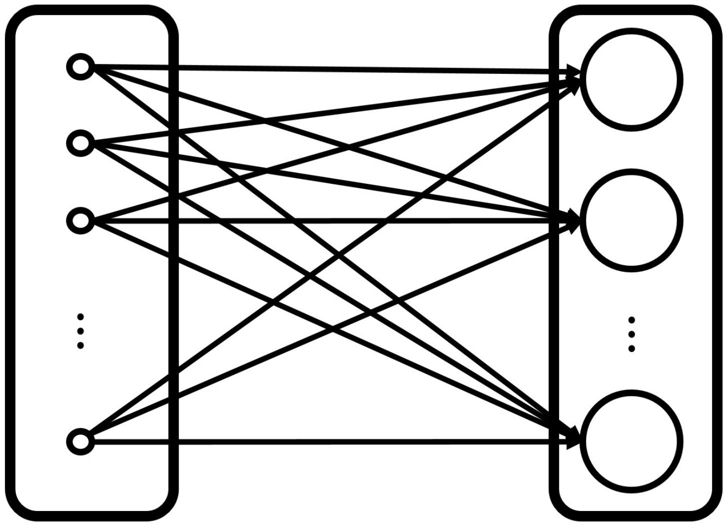
۵.۳ لایه کاملاً متصل

لایه کاملاً متصل[۱۲] که گاهی با نام لایه چگال[۱۳] نیز شناخته می‌شود، از ساختاری مشابه با ساختار شبکه‌های پرسپترون چند لایه بهره می‌برد. این لایه از دسته‌ای از واحدهای عصبی تشکیل شده است که هر یک از آن‌ها به صورت مستقل به جمع وزن‌دار ورودی‌های لایه می‌پردازند. لایه‌های کاملاً متصل در انتهای شبکه‌های کانولوشنی قرار می‌گیرند و وظیفه تولید خروجی نهایی را برای تصاویر ورودی بر عهده دارند. یک لایه کاملاً متصل در شکل ۴.۳ نشان داده شده است.

۶.۳ توابع فعالیت

توابع فعالیت[۱۴] توابع غیر خطی هستند که معمولاً پس از لایه‌های کانولوشنی یا لایه‌های کاملاً متصل مورد استفاده قرار می‌گیرند. دلیل استفاده از این لایه‌ها اضافه نمودن قابلیت یادگیری غیرخطی به شبکه است. معروفترین تابع فعالیتی که در شبکه‌های کانولوشنی مورد استفاده قرار می‌گیرد تابع فعالیت «واحد خطی یک‌سویه»[۱۵] یا $ReLU$ است [۹۰]. این تابع

[۱۲] Fully Connected
[۱۳] Dense
[۱۴] Activation Functions
[۱۵] Rectified Linear Unit



شکل ۴.۳: لایه‌های کاملاً متصل

نسبت به توابعی مانند سیگموئید[۱۶] دارای چندین مزیت است:

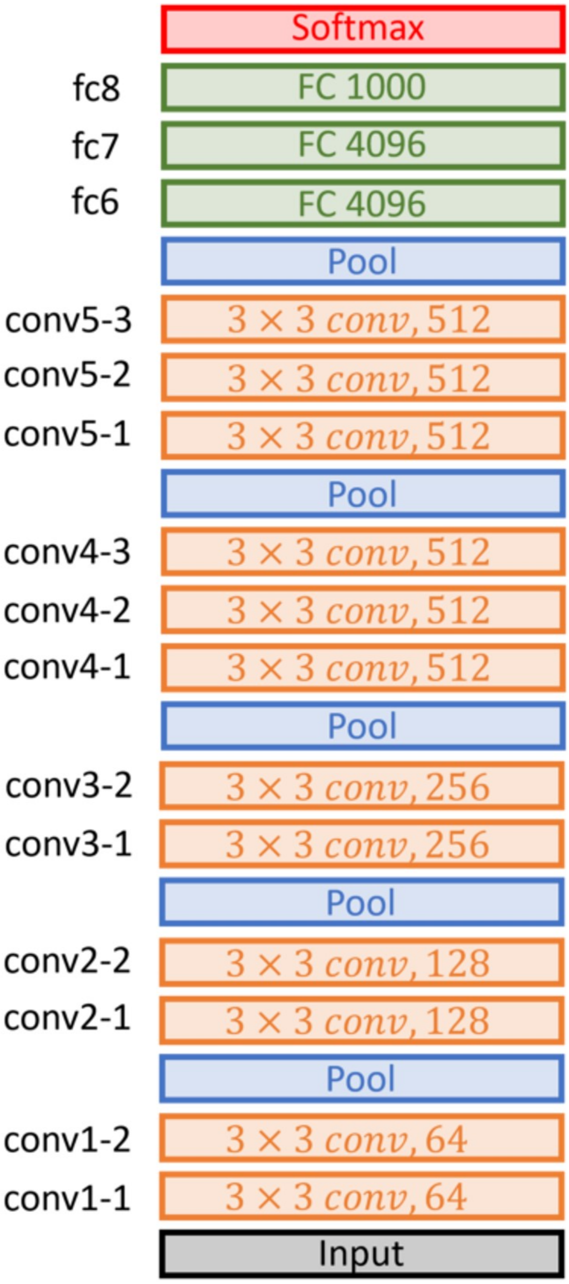
- این تابع برخلاف سیگموئید که با مشکل «ناپدید شدن گرادیان»[۱۷] رو به روست می‌تواند در شبکه‌هایی با لایه‌های بیشتر نیز به خوبی مقدار گرادیان را انتقال دهد.
- پیاده‌سازی آن آسان‌تر و سریع‌تر از سیگموئید است.
- شبکه‌هایی که از این تابع استفاده می‌کنند زودتر از شبکه‌های مشابه که از تابع سیگموئید استفاده می‌کنند همگرا می‌شوند [۹۱].

فرمول تابع سیگموئید در رابطه (۱.۳) و تابع $ReLU$ در رابطه (۲.۳) آمده است.

$$S(x) = \frac{1}{1 + e^{-\alpha x}} \tag{۱.۳}$$

$$R(x) = max(x, 0) \tag{۲.۳}$$

[۱۶] Sigmoid
[۱۷] Vanishing Gradient



شکل ۵.۳: مشخصات شبکه $VGG16$: برای لایه‌های کانولوشنی اندازه فیلترها و تعداد آن‌ها ذکر شده است. برای لایه‌های کاملاً‌ـ‌متصل نیز تعداد نرون‌ها برای هر لایه ذکر شده است.

۷.۳ ساختار کلی شبکه‌های کانولوشنی

ساختار کلی شبکه‌های کانولوشنی برای کلاسه‌بندی یا رگرسیون معمولاً به این صورت است که در ابتدا چندین لایه کانولوشنی قرار داده می‌شوند که در میان آن‌ها از لایه‌های ادغام

بیشینه استفاده می‌شود. به عنوان یک روند کلی نیز در لایه‌های متوالی ابعاد لایه‌ها کاهش پیدا کرده و تعداد کانال‌های هر لایه افزایش پیدا می‌کند. در [۲۱] نشان داده شده است که لایه‌های متوالی کانولوشنی با فیلترهای ۳ × ۳ قابلیت شبیه‌سازی استفاده از فیلترهای با اندازه دلخواه را دارند؛ بنابراین معمولاً در شبکه‌های کانولوشنی تنها از فیلترهای ۳ × ۳ استفاده می‌شود. شکل ۵.۳ جزئیات شبکه $VGG16$ را نشان می‌دهد.

این شبکه در کلاسه‌بندی و مکان‌یابی اشیای «چالش بازشناسی بصری در ابعاد بزرگ»[۱۸] در سال ۲۰۱۴ [۹۲] به ترتیب رتبه‌های اول و دوم را کسب کرد. شبکه‌های پیش از VGG عموماً تنها روی ابعاد پنجره‌ها و یا گام‌های حرکت پنجره‌ها متمرکز شده بودند، امّا این شبکه موضوع مهم عمق شبکه را مطرح کرد. این شبکه نشان داد که می‌توان با استفاده از فیلترهای کوچک‌تر متوالی تعداد پارامترهای شبکه را کاهش داد؛ همچنین می‌توان بدون افزایش تعداد پارامترها عمق شبکه را افزایش داد که باعث می‌شود توابع فعالیت بیشتری در شبکه مورد استفاده قرار گیرند و قدرت تفکیک شبکه بیشتر شود.

ورودی این شبکه تصاویر رنگی هستند؛ لایه‌های کانولوشنی این شبکه به گروه‌هایی تقسیم شده‌اند که بین آن‌ها لایه‌های ادغام بیشینه با اندازه پنجره و گام ۲ قرار دارند. هر یک از این گروه‌ها نیز به چند زیرگروه تقسیم می‌شوند که مابین آن‌ها توابع فعالیت $ReLU$ قرار داده شده‌اند. در این شبکه با هر بار اعمال شدن لایه‌های ادغام ابعاد تصویر نصف می‌شوند؛ امّا تعداد کانال‌ها دو برابر می‌شود؛ به طوری که لایه کانولوشنی اول دارای ۶۴ فیلتر بوده و آخرین لایه کانولوشنی ۵۱۲ فیلتر دارد. در نهایت سه لایه کاملاً_متصل قرار دارند که دو لایه اول ۴۰۹۶ نرون دارند و لایه آخر ۱۰۰۰ نرون دارد؛ دلیل استفاده از ۱۰۰۰ نرون در لایه نهایی تعداد کلاس‌های مجموعه داده‌ای است که شبکه با استفاده از آن آموزش داده شده است. این شبکه با مجموعه‌داده‌ای که حاوی بیش از یک میلیون تصویر و ۱۰۰۰ کلاس است آموزش داده شده است.

[۱۸] Large Scale Visual Recognition Challange

۸.۳ جمع‌بندی

با توجه به اینکه دو روش از چهار روش ارائه شده در این رساله از یادگیری عمیق استفاده می‌کنند، شبکه‌های عصبی کانولوشنی که گونه‌ای از یادگیری عمیق هستند در این فصل مورد بررسی قرار گرفتند. این شبکه‌ها بر خلاف روش‌های پیش از خود، هم نحوه استخراج ویژگی از تصاویر ورودی را یاد می‌گیرند و هم نحوه کلاسه‌بندی آن‌ها را می‌آموزند و هر دوی این وظایف را به صورت همزمان انجام می‌دهند. همچنین لایه‌های مختلفی که در شبکه‌های کانولوشنی کاربرد دارند در این فصل بررسی شدند. از میان این لایه‌ها، لایه‌های کانولوشن، ادغام بیشینه، کاملاً ـ متصل و توابع فعالیت آن‌ها با توجه به کاربرد زیادشان مورد بررسی قرار گرفته و شرح داده شده‌اند.

فصل ۴

امضای رنگ لبه‌ها

۱.۴ مقدمه

از آن‌جا که متون موجود در تصاویر صحنه طبیعی بیشتر اوقات برای جلب توجه انسان مورد استفاده قرار می‌گیرند معمولا دارای لبه‌های نسبتا قوی و کنتراست مناسب نسبت به پس‌زمینه خود هستند. بر اساس مشاهدات ما، معمولا سه رنگ غالب در اطراف لبه‌های نواحی متنی وجود دارند: رنگ متن (پیش‌زمینه)، رنگ پس‌زمینه و رنگ مرز میان این دو بخش؛ رنگ بخش مرزی گاهی اوقات یک رنگ متفاوت از دو رنگ متنی و پس‌زمینه است و در پاره‌ای از مواقع نیز مخلوطی از این دو رنگ است.

ما به کمک رنگ‌های غالب اطراف لبه، امضای رنگ لبه را تعریف می‌کنیم. از آنجایی که امضای لبه در بیشتر اوقات برای پیکسل‌های یک خط از متن یکسان است، از این واقعیت استفاده کرده و نقاطی با امضای رنگ مشابه را گروه‌بندی می‌کنیم که هر یک از این گروه‌ها در ادامه روند الگوریتم به عنوان نواحی کاندیدای متنی شناخته می‌شوند. شکل ۱.۴ فلوچارت

روش پیشنهادی اول را نشان می‌دهد.

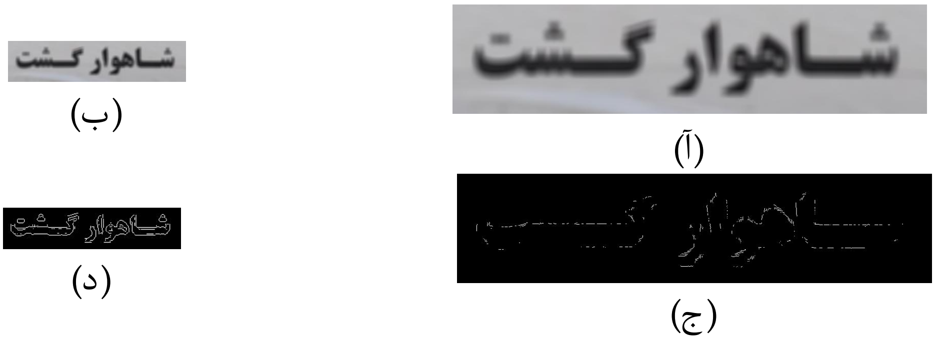
۲.۴ ایجاد هرمی از مقیاس‌های مختلف تصویر ورودی

در این مرحله از الگوریتم، هرمی از مقیاس‌های مختلف (۱, ۰/۸, ۰/۶, ۰/۴, ۰/۲) تصویر ورودی تشکیل می‌شود؛ بطوریکه مراحل بعدی الگوریتم به صورت جداگانه روی تمامی لایه‌های این هرم اعمال شده و در نهایت خروجی این هرم‌ها با هم ترکیب می‌شوند. دو دلیل برای استفاده از هرم تصویر وجود دارد:

- با توجه به اینکه متون موجود در تصاویر صحنه‌ی طبیعی دارای شکل‌ها و اندازه‌های مختلف هستند استفاده از هرم تصاویر باعث می‌شود تا الگوریتم بتواند اندازه‌ها و فونت‌های مختلف متون را مدیریت کرده و آن‌ها را شناسایی کند.

- مراحل بعدی الگوریتم از لبه‌ها برای شناسایی نواحی کاندیدای متنی استفاده می‌کنند. اما در تصاویری که دارای محوشدگی هستند شناسایی لبه‌ها به خوبی انجام نمی‌گیرد. دلیل این موضوع آن است که روش‌های استخراج لبه معمولا از فیلترهایی با ابعاد ثابت برای شناسایی لبه‌ها استفاده می‌کنند اما طبق مشاهدات ما با توجه به این که در تصاویر دارای محو شدگی عرض لبه‌ها از اندازه فیلترها بیشتر می‌شود، در نتیجه لبه‌ها در این مناطق به اندازه کافی قوی نیستند تا به راحتی شناسایی شوند. اما هنگامی که همان تصویر با مقیاس کوچکتری مورد استفاده قرار می‌گیرد عرض لبه‌ها نیز کاهش پیدا کرده و این باعث می‌شود که لبه‌ها قوی‌تر شده و راحت‌تر شناسایی شوند. این پدیده در شکل ۲.۴ نشان داده شده است. همانگونه که مشاهده می‌شود لبه‌های تصویر برای مقیاس اولیه به خوبی شناسایی نشده‌اند اما برای مقیاس کوچکتر دارای کیفیت خوبی هستند.

شکل ۱.۴: فلوچارت روش پیشنهادی اول



شکل ۲.۴: تاثیر کاهش ابعاد تصویر در کیفیت لبه‌های خروجی الگوریتم

۳.۴ استخراج لبه‌ها

در مرحله بعدی الگوریتم، لبه‌های تصویر ورودی با استفاده از روش سوبل [1] استخراج می‌شوند. استفاده از این روش به دو دلیل بوده است:

- روش سوبل در عین حال که برای بخش‌های متنی لبه‌های خوبی استخراج می‌کند، برای بخش‌های غیر متنی نیز لبه‌های کمتری استخراج می‌کند که روی کارایی الگوریتم در مراحل بعدی تاثیر مثبت دارد. این موضوع در شکل ۳.۴ نشان داده شده است. همانگونه که در شکل ۳.۴ج مشاهده می‌شود روش سوبل به خوبی لبه‌های متنی را استخراج کرده است و در بخش‌های غیر متنی نیز لبه‌های بسیار کمتری استخراج کرده است.

- دلیل دوم عدم استفاده از روش کنی [2] آن است که این روش در مناطق متنی نیز لبه‌های اضافی زیادی تولید می‌کند؛ این لبه‌های اضافی برای مراحل بعدی الگوریتم بار محاسباتی ایجاد می‌کنند. این موضوع در شکل ۴.۴ نیز نشان داده شده است. همانگونه که مشاهده می‌شود روش کنی لبه‌های ناخواسته بسیار زیادی ایجاد کرده است.

---

[1] Sobel
[2] Canny

(ج) (ب) (آ)

شکل ۳.۴: مقایسه تعداد پیکسل‌های لبه استخراج شده توسط روش‌های سوبل و کنی برای یک تصویر نمونه آ) تصویر ورودی ب) خروجی روش کنی ج) خروجی روش سوبل

۴.۴ هرس کردن پیکسل‌های غیر متنی

با توجه به اجزای متصل [۳] لبه‌های متنی دارای ویژگی‌های خاصی هستند در این بخش از الگوریتم از این ویژگی‌ها برای هرس کردن بخشی از لبه‌های غیر متنی استفاده می‌کنیم. با توجه به اینکه لبه‌های متنی بسیار بلند و یا بسیار کوتاه نیستند، بنابراین نسبت طول به ارتفاع مستطیل محیط [۴] به آن‌ها معمولا در یک بازه مشخص قرار دارد. همچنین با توجه به اینکه لبه‌های بخش‌های متنی مانند شاخ و برگ درختان دارای لبه‌های بسیار پیچیده (مانند شاخ و برگ درختان) و یا بسیار ساده (مانند قاب دور تابلوها) نیستند، برای یک جزء متصل لبه‌های متنی، نسبت تعداد پیکسل‌های متنی به مساحت مستطیل احاطه کننده آن‌ها در یک بازه مشخص قرار دارد. در این گام الگوریتم تعدادی قانون برای هرس کردن اجزای غیر متنی مورد استفاده قرار می‌گیرد تا بار پردازشی الگوریتم در مراحل بعدی کاهش پیدا کند.قوانین مورد استفاده برای هرس کردن اجزای لبه غیر متنی در ادامه آمده است:

$$\alpha_1 \leq \frac{w}{h} \leq \alpha_2 \qquad (۱.۴)$$

$$\alpha_3 \leq \frac{area}{w \times h} \leq \alpha_4 \qquad (۲.۴)$$

[۳] Connected Components
[۴] Bounding Box

(آ)

(ب)

(ج)

(د)

شکل ۴.۴: مقایسه عملکرد روش‌های سوبل و کنی در مناطق متنی (لبه‌ها با رنگ مشکی نشان داده شده‌اند) آ) تصویر ورودی ب) برشی از تصویر ورودی ج) خروجی روش کنی د) خروجی روش سوبل

$$\frac{area}{ma} \leq \alpha_{5} \qquad (۳.۴)$$

$$\frac{f}{ma} \leq \alpha_{6} \qquad (۴.۴)$$

که در آن w و h به ترتیب طول و ارتفاع مستطیل محیط به جزء متصل، $area$ تعداد پیکسل‌های آن، ma طول قطر اصلی آن و f نسبت میان فاصله کانونی بیضی با ممان دوم[۵] یکسان با جزء متصل، به طول محور اصلی [۶] آن است.

محدودیت‌های (۱.۴) و (۴.۴) به بررسی این نکته می‌پردازند که آیا طول و درازای جزء مورد نظر در محدوده متصور برای اجزای متنی قرار دارد. محدودیت (۴.۴) همیشه در بازه $[0, 1]$ قرار دارد مقادیر بزرگتر آن نشان دهنده کشیدگی بیشتر ناحیه است. در واقع محدودیت‌های (۱.۴) و (۴.۴) تقریباعملکرد مشابهی دارند؛ دلیل استفاده از دو محدودیت آن است که محدودیت اول بسیار سریع‌تر و آسان‌تر از دومی محاسبه می‌شود بنابراین از آن در مرحله اول استفاده می‌شود اما با توجه به اینکه قادر نیست نواحی خطی مورب را تشخیص دهد، در مرحله نهایی محدودیت (۴.۴) با پیچیدگی بیشتر مورد استفاده قرار می‌گیرد.

معمولاً تعداد زیادی پیکسل لبه در یک ناحیه متنی قرار نمی‌گیرند و چگالی زیاد پیکسل در یک ناحیه می‌تواند نشان‌دهنده غیر متنی بودن آن ناحیه باشد. محدودیت (۲.۴) به بررسی این موضوع پرداخته و نواحی را که از نظر تعداد پیکسل‌های لبه بسیار شلوغ یا خلوت هستند هرس می‌کند. در نهایت محدودیت (۳.۴) نیز به بررسی این نکته در نواحی می‌پردازد که آیا ناحیه مورد نظر دارای پیچ و خم‌های بسیار زیاد است یا خیر. با توجه به اینکه نواحی متنی بسیار پیچیده نیستند بنابراین این معیار نیز برای نواحی متنی از حد مشخصی تجاوز

[۵] Second Moment
[۶] Major Axis

نخواهد کرد. ترتیب اعمال شدن محدودیت‌ها نیز به همان ترتیب قرار گرفتن آن‌ها در متن است. مقادیر پارامترهای موجود در محدودیت‌ها نیز به صورت تجربی از روی تصاویر آموزشی محاسبه می‌شود. تاثیر هر یک از این محدودیت‌ها در شکل ۴.۴ برای یک تصویر نمونه نشان داده شده است.

۵.۴ امضای رنگ لبه‌ها

برای هر یک از لبه‌های متنی که در مرحله قبل حذف نشده باشند یک توصیف‌گر در این مرحله استخراج می‌شود که ما آن را امضای رنگ [۷] نام‌گذاری کرده‌ایم.

تعریف ۱.۵.۴. مجموعه رنگ‌های غالب در همسایگی یک پیکسل لبه در تصویری که لبه‌ها از آن به دست آمده‌اند امضای رنگ آن پیکسل نامیده می‌شود.

برای استخراج امضای پیکسل‌های لبه، الگوریتم انتقال میانگین [۸] در یک محدوده $n \times n$ در اطراف پیکسل لبه برای استخراج رنگ‌های غالب مورد استفاده قرار گرفت که در ادامه شرح داده می‌شود. این الگوریتم از قبل نیاز به دانستن تعداد خوشه‌ها نداشته و خروجی آن تعدادی مرکز خوشه است. هر کدام از این مراکز خوشه یکی از رنگ‌های مربوط به امضای رنگ است.

۱.۵.۴ الگوریتم انتقال میانگین

انتقال میانگین یک الگوریتم جستجوی مد [۹] غیرپارامتری [۱۰] است که می‌تواند به عنوان یک الگوریتم خوشه‌بندی نیز مورد استفاده قرار گیرد [۹۳]. برخلاف الگوریتم‌هایی مانند k-means، در این الگوریتم لازم نیست که تعداد خوشه‌ها از قبل مشخص شوند. اگر n نمونه داده $x_i, i = 1 \ldots n$ در فضای d بعدی R^d داشته باشیم، تخمین تابع چگالی احتمال با

---

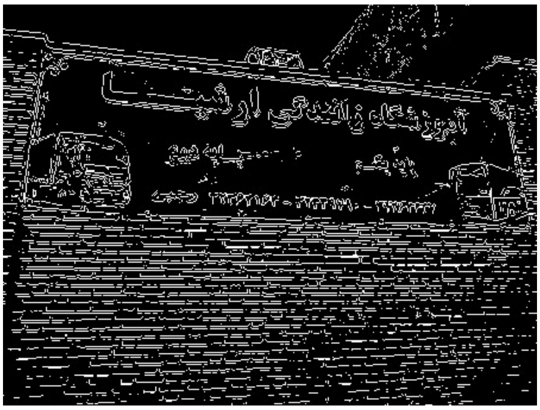
[۷] Color Signatures

[۸] Mean Shift

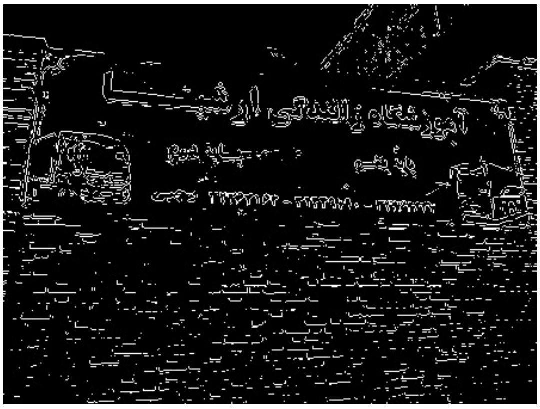
[۹] Mode Seeking

[۱۰] Non Parametric

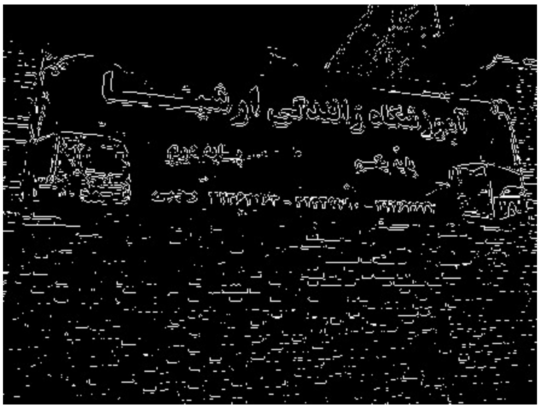
(آ)



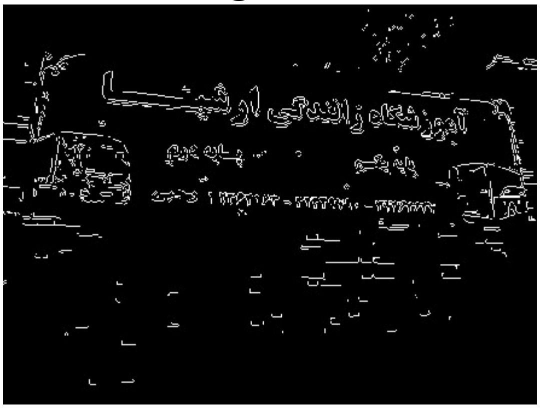
(ب)



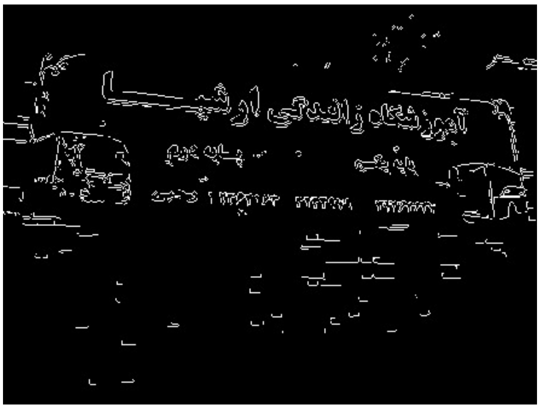
(ج)



(د)



(ه)



(و)

شکل ۵.۴: تاثیر استفاده از قواعد هندسی برای هرس کردن نواحی غیر متنی، آ) تصویر ورودی ب) لبه‌های تصویر ورودی ج) تاثیر اعمال محدودیت (۱.۴) د) تاثیر اعمال محدویت (۲.۴) ه) تاثیر اعمال محدودیت (۳.۴) و) تاثیر اعمال محدودیت (۴.۴)

هسته[۱۱] K و عرض[۱۲] h به صورت زیر انجام می‌گیرد [۹۴]:

$$f(x) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \tag{۵.۴}$$

و هسته‌های متقارن به صورت زیر تعریف می‌شوند [۹۴]:

$$K(x) = c_{k,d} k\left(||x||^2\right) \tag{۶.۴}$$

که در آن $c_{k,d}$ یک عامل نرمال‌سازی است. با فرض مشتق‌پذیر بودن K، مدهای تابع f در صفرهای گرادیان آن به صورت زیر قرار دارند [۹۴]:

$$\begin{aligned} \nabla f(x) &= \left(\frac{2c_{k,d}}{nh^{d+2}}\right) \sum_{i=1}^{n} (x_i - x) g(\left(\left|\left|\frac{x - x_i}{h}\right|\right|^2\right) \\ &= \left(\frac{2c_{k,d}}{nh^{d+2}}\right) \left(\sum_{i=1}^{n} g\left(\left|\left|\frac{x - x_i}{h}\right|\right|^2\right)\right) \left(\frac{\sum_{i=1}^{n} x_i g(\left(||\frac{x-x_i}{h}||^2\right)}{\sum_{i=1}^{n} g(\left(||\frac{x-x_i}{h}||^2\right)}\right) \end{aligned} \tag{۷.۴}$$

که در آن $g(s) = -k'(s)$ است. عبارت اول رابطه (۷.۴) با تعریف $G(x) = c_{k,d} g(x^2)$ به عنوان هسته، کسری از هسته خواهد بود و عبارت دوم، بردار انتقال میانگین نامیده می‌شود [۹۴]:

$$m_h(x) = \frac{\sum_{i=1}^{n} x_i g\left(||\frac{x-x_i}{h}||^2\right)}{\sum_{i=1}^{n} g\left(||\frac{x-x_i}{h}||^2\right)} - x \tag{۸.۴}$$

بردار انتقال میانگین همواره به جهت افزایش حداکثری تابع چگالی اشاره می‌کند [۹۵]؛ با استفاده از این واقعیت، الگوریتم انتقال میانگین با استفاده از اعمال مکرر رابطه زیر تحقق می‌یابد:

---

[۱۱] Kernel

[۱۲] Bandwidth

$$x^{t+1} = x^t + m_h(x^t) \quad (۹.۴)$$

که در آن x^t جواب مسئله در تکرار جاری را نشان می‌دهد. در نهایت x به مد تابع چگالی همگرا خواهد شد که در آنجا گرادیان تابع چگالی با صفر برابر است. الگوریتم «خوشه‌بندی انتقال میانگین» یک کاربرد عملی از الگوریتم «جستجوی مد انتقال میانگین» است. این الگوریتم کار خود را از چندین نقطه تصادفی شروع می‌کند. الگوریتم انتقال میانگین به هر یک از این نقاط اعمال می‌شود که در نهایت هر یک از نقاط به یک مد هم‌گرا می‌شوند؛ این مدها به عنوان مرکز خوشه در نظر گرفته می‌شوند و تعداد آن‌ها تعداد خوشه‌ها را نشان می‌دهد.

۶.۴ خوشه‌بندی نقاط لبه با استفاده از امضای رنگ

در این مرحله از الگوریتم، پیکسل‌های لبه با توجه به امضای رنگ آن‌ها و محل قرار گرفتن آن‌ها با استفاده از الگوریتم اتصال تکی [۱۳] خوشه‌بندی می‌شوند. الگوریتم اتصال تکی یک الگوریتم خوشه‌بندی سلسله‌مراتبی است؛ این الگوریتم در ابتدا با تعداد خوشه برابر با نقاط داده ورودی کار را شروع می‌کند. در هر مرحله از این الگوریتم، دو خوشه که از بقیه خوشه‌ها به هم نزدیک‌تر هستند با هم ادغام می‌شوند. ما فاصله بین دو خوشه را فاصله نزدیکترین نقاط آن دو خوشه نسبت به هم تعریف می‌کنیم. با تعریف یک معیار فاصله مناسب، این الگوریتم قادر است تا نقاط لبه را به صورت زنجیروار به هم متصل کرده و مناطق کاندیدای متنی را تشکیل دهد. ما فاصله بین دو نقطه p و q را به صورت زیر تعریف می‌کنیم:

$$d(p,q) = \sum_{i=1}^{c_p} \min_{j=1\dots c_q} d_{ij}(p,q) + \beta \times ||pos_p - pos_q|| \quad (۱۰.۴)$$

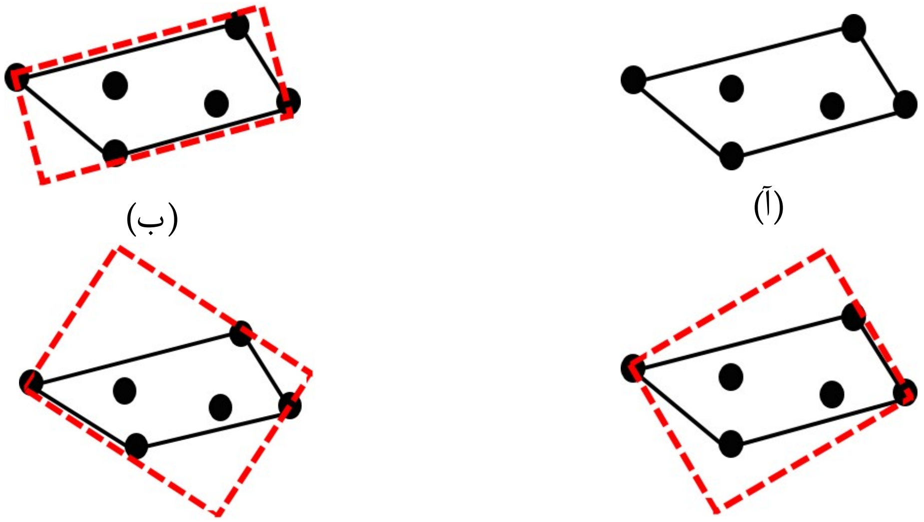
---

[۱۳] Single Linkage

$$d_{ij}(p,q) = \sum_{k\in\{r,g,b\}} |k_{p_i} - k_{q_i}| \tag{۱۱.۴}$$

$$||pos_p - pos_q|| = |p_x - q_x| + |p_y - q_y|^۲ \tag{۱۲.۴}$$

که در آن c_x نشان‌دهنده تعداد رنگ‌ها برای نقطه x، pos_x یک بردار دو بعدی نشان‌دهنده موقعیت مکانی نقطه x، پارامتر β اهمیت نسبی امضای رنگ و فاصله مکانی را کنترل می‌کند، کانال‌های رنگی توسط r، g و b نشان داده شده‌اند و در نهایت k_x نشان دهنده کانال k خوشه x است. همانگونه که از رابطه (۱۰.۴) نیز بر می‌آید، تعداد خوشه‌ها یا به عبارت دیگر رنگ‌های به دست آمده برای نقاط لبه متفاوت ممکن است با هم فرق داشته باشد؛ همچنین ترتیب خوشه‌ها نیز می‌تواند متفاوت باشد، بنابراین برای به دست آوردن فاصله دو نقطه عملگر min مورد استفاده قرار گرفت. همچنین عملگر $|.|$ نشان دهنده قدر مطلق است. توان ۲ در رابطه (۱۲.۴) به این دلیل اعمال شده است تا برای فواصل خطوط متنی که در مجاورت هم قرار گرفته‌اند مقدار زیادی به دست آمده و از ادغام شدن آن‌ها جلوگیری شود. هم رنگ و هم فاصله مکانی در محاسبه فاصله بین دو نقطه نقش ایفا می‌کنند، این موضوع در رابطه (۱۰.۴) نیز در نظر گرفته شده و فاصله نهایی ترکیبی از این دو معیار خواهد بود. اما در صورتی که فاصله امضای رنگ دو نقطه با هم از حد از پیش تعیین شده‌ای تجاور کند عمل ادغام صورت نمی‌گیرد. ادغام خوشه‌ها نیز تا رسیدن به یک فاصله خوشه معین ادامه پیدا می‌کند. در نهایت خوشه‌های بسیار کوچک که شامل کمتر از ۲۰ پیکسل باشند کنار گذاشته می‌شوند. خوشه‌های باقی مناطق کاندیدای متنی را تشکیل می‌دهند به این صورت که هر خوشه یک کاندیدای ناحیه متنی به شمار می‌آید.



شکل ۶.۴: نحوه بررسی کاندیداهای مختلف برای به دست آوردن مستطیل محیط کمینه

۷.۴ استخراج ناحیه

پس از یافتن مناطق کاندیدای متنی، آن‌ها از تصویر ورودی استخراج می‌شوند. در گام اول از این مرحله یک مستطیل محیط کمینه برای نقاط لبه هر یک از نواحی کاندید به دست می‌آید. با توجه به اینکه همیشه یکی از اضلاع مستطیل محیط کمینه هم راستا با یکی از اضلاع پوسته محدب [۱۴] مربوط به مناطق مورد نظر است [۹۶]، بنابراین می‌توان با به دست آوردن غلاف محدب و بررسی تمامی مستطیل‌های محیط که این شرایط را دارند در نهایت مستطیل کمینه را به دست آورد. این موضوع در شکل ۶.۴ نشان داده شده است.

پس از یافتن مستطیل محیط کمینه، در مرحله بعدی زاویه آن را با افق به دست آورده و به شکلی آن را می‌چرخانیم که ضلع بلندتر آن با افق هم راستا شود؛ در نهایت ناحیه مذکور را به عنوان ناحیه کاندید از تصویر استخراج می‌کنیم. شکل ۷.۴ این روند را توصیف می‌کند.

[۱۴] Convex Hull

(ب)

(آ)

شکل ۷.۴: نحوه استخراج ناحیه کاندید از تصویر آ) یافتن زاویه ناحیه با افق ب) ناحیه استخراج شده

۸.۴ کلاسه‌بندی

در این بخش از الگوریتم نواحی کاندیدای استخراج شده از تصویر با استفاده از یک کلاسه‌بندی آبشاری[۱۵] به متن یا غیر متن کلاسه‌بندی می‌شوند. این کلاسه‌بندها در زمینه‌های مختلفی از جمله تشخیص اشیا یا چهره مورد استفاده قرار گرفته و کارایی خود را به اثبات رسانده‌اند [۹۷، ۹۸]. این کلاسه‌بندها سریع بوده و نتایج مناسبی حتی برای مسائل سخت تولید می‌کنند. نحوه کار کلاسه‌بندهای آبشاری به این شکل است که به جای آن که تمامی ویژگی‌های مورد استفاده در کلاسه‌بندی یکجا و به صورت دفعی مورد استفاده قرار گیرند به صورت مرحله‌ای و در گام‌های متوالی استفاده می‌شوند. هر گام الگوریتم ویژگی‌های متفاوتی برای کلاسه‌بندی داده‌ها مورد استفاده قرار می‌دهد و بخشی از نمونه‌های کلاس منفی (غیر متن) را هرس می‌کند. مزیت این رویکرد را می‌توان در دو مورد بیان کرد:

- این نوع کلاسه‌بندها همه ویژگی‌ها را یکجا استخراج نمی‌کنند و در هر مرحله از کلاسه‌بند تعدادی از نواحی با برچسب منفی کنار گذاشته می‌شوند؛ بنابراین ویژگی‌ها تنها زمانی استخراج می‌شوند که به آن‌ها نیاز وجود داشته باشد. با این روش می‌توان ویژگی‌هایی را که به لحاظ محاسباتی پیچیده هستند در انتهای آبشار قرار داد تا سرعت کلاسه‌بند افزایش پیدا کند.

---
[۱۵] Cascade

- مزیت بعدی استفاده از این نوع کلاسه‌بندها دوری جستن از مشکل ابعاد[۱۶] است. در این نوع کلاسه‌بندها در هر مرحله می‌توان ویژگی‌های متفاوتی مورد استفاده قرار داد بدون آن که نگران مشکل ابعاد باشیم. البته در مواردی که در هر مرحله ویژگی جدید به ویژگی‌های قبلی اضافه شود این مشکل همچنان پا برجا خواهد بود.

- در نهایت مزیت دیگر استفاده از این نوع کلاسه‌بندها آن است که کنار هم قرار دادن ویژگی‌ها باعث افزایش تعداد پارامترهای کلاسه‌بند می‌شود که به نوبه خود باعث افزایش نیاز به داده‌های بیشتر برای آموزش کلاسه‌بند می‌شود. در مواردی که داده‌های موجود زیاد نباشند این مسئله مشکل‌ساز است.

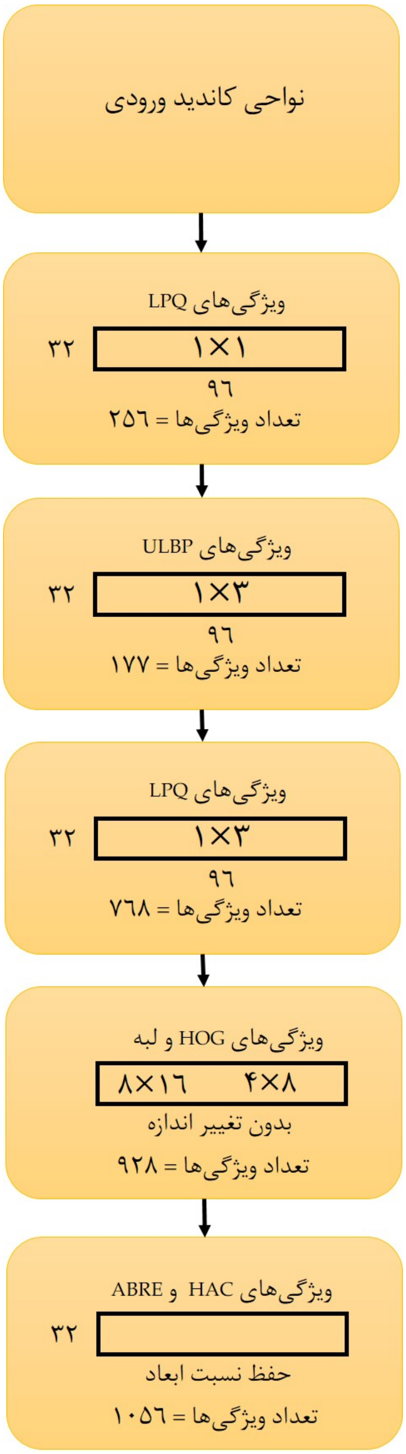
در هر مرحله از کلاسه‌بند آبشاری پیشنهادی، یک کلاسه‌بند شبکه عصبی پرسپترون چند لایه مورد استفاده قرار می‌گیرد. ویژگی‌های ورودی و همچنین ساختار هر یک از این کلاسه‌بندها با هم متفاوت است. از آنجایی که در هر مرحله از کلاسه‌بند آبشاری کار نسبت به مراحل قبلی سخت‌تر می‌شود، بنابراین باید ویژگی‌ها و کلاسه‌بند متناسب هر مرحله مورد استفاده قرار گیرد. در نتیجه در مراحل متوالی کلاسه‌بند آبشاری، تعداد نرون‌های لایه میانی را افزایش می‌دهیم تا برای کلاسه‌بندی نمونه‌های سخت‌تر مناسب شود. برای بخش اول کلاسه‌بند یک شبکه عصبی با ۱۰ نرون لایه میانی مورد استفاده قرار می‌گیرد و در مراحل بعدی در هر مرحله ۱۰ نرون به نرون‌های لایه قبلی اضافه می‌شود. شکل ۸.۴ برخی از جزئیات ویژگی‌های مورد استفاده در هر یک از لایه‌های کلاسه‌بند آبشاری را نشان می‌دهد.

۱.۸.۴ سه لایه ابتدایی

در سه لایه ابتدایی، ناحیه‌ای که قرار است کلاسه‌بندی شود به ابعاد ثابت ۹۶ × ۳۲ تغییر اندازه داده می‌شود. در لایه اول ویژگی‌های چندی‌سازی فاز محلی[۱۷] مورد استفاده قرار می‌گیرد [۹۹]. این ویژگی‌ها با چندی‌سازی بخش فاز تبدیل فوریه گسسته به دست می‌آیند.

[۱۶] Curse of Dimensionality

[۱۷] Local Phase Quantization



شکل ۸.۴: جزئیات ویژگی‌های مورد استفاده در لایه‌های مختلف کلاسه‌بند آبشاری

این ویژگی در مقابل کاربردهای مختلفی مورد استفاده قرار گرفته و کارایی مناسبی از خود نشان داده‌اند. این ویژگی هم دارای دقت مناسب است و هم سرعت خوبی دارد و در مقابل محو شدگی تصاویر هم مقاوم است که آن را تبدیل به یک گزینه مناسب برای مرحله اول کلاسه‌بند آبشاری می‌نماید.

در لایه بعدی ویژگی‌های الگوی دودویی محلی یکنواخت[۱۸] [۱۰۰] به کار گرفته شد. در ابتدا ناحیه ورودی به ۳ بخش غیر هم‌پوشان تقسیم شد. اندازه هر یک از این بخش‌ها ۳۲ × ۳۲ است. در ادامه ویژگی‌های الگوی دودویی محلی از هر یک از این نواحی استخراج می‌شوند. در حالی که ویژگی‌های مرحله اول ناحیه را به صورت کلی مورد بررسی قرار می‌دهند، ویژگی‌های مورد استفاده در این بخش حاوی اطلاعات بیشتری در مورد ناحیه مورد نظر هستند.

در لایه بعدی نیز ناحیه ورودی به ۳ بخش غیر هم‌پوشان ۳۲ × ۳۲ تقسیم شده و ویژگی‌های چندی‌سازی فازی محلی از آن‌ها استخراج می‌شود. در اینجا نیز با توجه به افزایش تعداد سلول‌های مورد استفاده اطلاعات بیشتری از ناحیه مورد نظر استخراج می‌شود.

۲.۸.۴ لایه چهارم

برای لایه چهارم، ویژگی‌های هیستوگرام گرادیان جهت‌دار [۱۹] و یک نسخه جدید از ویژگی‌های معرفی شده در [۴۱] استفاده می‌شوند. ویژگی‌های مورد استفاده در این مرحله پیچیده‌تر از ویژگی‌های مراحل قبلی بوده و بنابراین برای افزایش سرعت کلاسه‌بند، در مراحل انتهایی آبشار قرار گرفته‌اند. ویژگی HOG بدون هم‌پوشانی و از بخش‌هایی با ابعاد ۸ × ۴ استخراج می‌شوند. برای استخراج ویژگی‌های سری دوم مربوط به این لایه، که ما آن‌ها را ویژگی‌های لبه می‌نامیم، در ابتدا اندازه گرادیان و لبه‌های ناحیه از کانال شدت روشنایی فضای $L*a*b$ [۱۰۱] به دلیل کنتراست بهتر آن به دست می‌آیند. در مرحله بعد ناحیه کاندیدا به سلول‌های غیر هم‌پوشان ۱۶ × ۸ تقسیم می‌شود. مجموعه ویژگی لبه ابتدایی که استخراج می‌شوند

[۱۸] Uniform Local Binary Pattern

[۱۹] Histogram of Oriented Gradient

میزان همگنی ناحیه کاندیدا را مورد بررسی قرار می‌دهد. این مجموعه ویژگی به صورت زیر تعریف می‌شود:

$$f_{۱}^{ij} = \frac{avg(G_e^{ij})}{std(G_e) + \varepsilon} \tag{۱۳.۴}$$

که در آن $\mu(G_e^{ij})$ مقدار میانگین اندازه گرادیان را در محل پیکسل‌های لبه برای سلول سطر iام و ستون jام نشان می‌دهد؛ همچنین $\sigma(G_e)$ مقدار انحراف استاندارد اندازه گرادیان را در محل سلول‌های لبه برای کل ناحیه کاندید نشان می‌دهد.

دو مجموعه ویژگی بعدی روی انعکاس‌های[۲۰] افقی و عمودی پیکسل‌های لبه به صورت زیر محاسبه می‌شوند:

$$f_{۲}^{ij} = \frac{num(v_{prj} \neq ۰) + num(h_{prj} \neq ۰)}{len(v_{prj}) + len(h_{prj})} \tag{۱۴.۴}$$

$$f_{۳}^{ij} = \frac{num(v_{prj} = \alpha k) + num(h_{prj} = \alpha k)}{len(v_{prj}) + len(h_{prj})}, \alpha = ۲, k = ۱, ۲, \ldots \tag{۱۵.۴}$$

$$f_{۴}^{ij} = \frac{num(v_{prj} = \alpha k) + num(h_{prj} = \alpha k)}{len(v_{prj}) + len(h_{prj})}, \alpha = ۳, k = ۱, ۲, \ldots \tag{۱۶.۴}$$

$$f_{۵}^{ij} = \frac{num(v_{prj} = \alpha k) + num(h_{prj} = \alpha k)}{len(v_{prj}) + len(h_{prj})}, \alpha = ۵, k = ۱, ۲, \ldots \tag{۱۷.۴}$$

[۲۰] Prjections

که در آن $vprj$ و $hprj$ به ترتیب انعکاس‌های عمودی و افقی بوده و $\#(.)$ تعداد عناصری از بردار ورودی را باز می‌گرداند که شرط ورودی را ارضا می‌کنند. همچنین len تعداد عناصر بردار ورودی خود را نشان می‌دهد.

مجموعه ویژگی‌های f_2 تعداد عناصر غیر صفر را در سلول‌ها می‌شمارند. همچنین مجموعه ویژگی‌های فرمول‌های (۱۵.۴) تا (۱۷.۴) تعداد ضرایب ۲، ۳ و ۵ را در میان انعکاس‌های عمودی و افقی شمارش می‌کنند. این مجموعه ویژگی‌ها سعی بر این دارند که نحوه توزیع شدن نقاط لبه را در بخش‌های مختلف نواحی کاندید مورد بررسی قرار دهند. مقادیر صفر در انعکاس‌های عمودی و افقی در تمامی مجموعه ویژگی‌های (۱۵.۴) تا (۱۷.۴) شمرده می‌شوند؛ بنابراین برای آنکه به صورت جداگانه نیز این شمارش‌ها انجام شود و میان مقادیر صفر و غیر صفر تفاوت قائل شویم، مجموعه ویژگی‌های f_2 مستقل از بقیه موارد تعریف شدند. همچنین این مجموعه ویژگی‌های شکل کلی انعکاس‌ها را نیز به دست می‌دهند. از آنجایی که برای محاسبه این مجموعه ویژگی‌ها ناحیه ورودی تغییر اندازه داده نمی‌شود بنابراین هر عددی ممکن است در خروجی آن‌ها ظاهر شود. برای پرهیز از این حالت و تولید اعداد بزرگ، به جای خود عدد تعداد ضرایب ۲، ۳ و ۵ شمرده می‌شوند. این اعداد به صورت تجربی بهترین نتایج را تولید کرده‌اند.

دلیل عدم تغییر اندازه در این مرحله جلوگیری از حذف اطلاعات مفیدی است که در صورت تغییر اندازه نواحی، که در بیشتر موارد نیز ناحیه کوچکتر می‌شود، از بین می‌روند. برخی مواقع نیز تغییر اندازه سبب تغییر نسبت طول به عرض تصویر می‌شود که به نوبه خود ساختارهای داخلی تصویر را تغییر داده و سبب به اشتباه افتادن کلاسه‌بند می‌شود.

۳.۸.۴ لایه پنجم

برای آخرین مرحله از کلاسه‌بند آبشاری، ناحیه ورودی با حفظ نسبت طول به عرض، به یک پنجره با ارتفاع ۳۲ پیکسل تغییر اندازه داده می‌شود. در ادامه ویژگی‌های $DFDL$ [۲۱] [۱۰۲]برای استخراج ویژگی از این ناحیه به کار گرفته می‌شوند. در هنگام آموزش، دو

[۲۱] Discrminative Feature-oriented Dictionary Learning

دیکشنری متفاوت برای داده‌های متنی و غیر متنی با استفاده ویژگی‌های $DSIFT$ [۲۲] یاد گرفته می‌شوند [۱۰۳]. در زمان آزمون، در ابتدا ویژگی‌های $SIFT$ از ناحیه ورودی استخراج می‌شوند. در ادامه، یک بار با استفاده از دیکشنری متنی و یک بار با استفاده از دیکشنری غیر متنی ویژگی‌های ناحیه ورودی به صورت تنک [۲۳] بازسازی می‌شود.

برای بازسازی هر یک از نواحی $SIFT$ استخراج شده با استفاده از یک دیکشتری، برای هر یک از الفبای [۲۴] دیکشنری یک وزن پیدا می‌شود. مجموع این وزن‌ها برای هر یک الفبا و هر یک از نواحی $SIFT$ یک جزء از بردار ویژگی اول این لایه است. ما این مجموعه ویژگی را هیستوگرام ضرایب الفبا [۲۵] (HAC) می‌نامیم. اگر فرض کنیم که d_i بردار ویژگی $SIFT$ استخراج شده iام باشد، در این صورت می‌توان آن را به صورت زیر هم توسط الفبای دیکشنری‌های متنی a_j^t و هم غیر متنی a_j^{nt} به صورت زیر بازسازی کرد:

$$\overrightarrow{d_i} \approx \overrightarrow{\hat{d_i}} = \sum_{j=1}^{n} \gamma_{ij}^t \overrightarrow{a_j^t} \quad (18.4)$$

$$\overrightarrow{d_i} \approx \overrightarrow{\hat{d_i}} = \sum_{j=1}^{n} \gamma_{ij}^{nt} \overrightarrow{a_j^{nt}} \quad (19.4)$$

در این صورت اگر تعداد الفبای دیکشنری‌ها را n و تعداد نواحی $SIFT$ استخراج شده را m در نظر بگیریم، ویژگی‌های HAC به صورت زیر به دست می‌آیند:

$$HAC = \{h_j | h_j = \sum_{i=1}^{m} \gamma_{ij}^t, j = 1 \dots n\} \cup \{h_k | h_k = \sum_{k=1}^{m} \gamma_{kj}^{nt}, k = 1 \dots n\} \quad (20.4)$$

[۲۲] Dense Scale Invariant Feature Transform

[۲۳] Sparse

[۲۴] Atom

[۲۵] Histogram of Atom Coefficients

ویژگی‌های بعدی که در این بخش استخراج می‌شوند مربوط به خطای هر یک از دیکشنری‌های متنی و غیر متنی برای بازسازی ناحیه ویژگی ورودی هستند. این ویژگی‌ها، که ما آن‌ها را خطای بازسازی ناحیه مطلق [۲۶] (ABRE) می‌نامیم، به صورت زیر محاسبه می‌شوند:

$$\overrightarrow{ABRE} = \left\{ \sum_{i=1}^{m} \left(\overrightarrow{d_i} - \sum_{j=1}^{n} \gamma_{ij}^{t} \overrightarrow{a_j^{t}} \right) \right\} \bigcup \left\{ \sum_{i=1}^{m} \left(\overrightarrow{d_i} - \sum_{j=1}^{n} \gamma_{ij}^{nt} \overrightarrow{a_j^{nt}} \right) \right\} \qquad (۲۱.۴)$$

این ویژگی‌ها سعی در بررسی این نکته دارند که دیکشنری‌های متنی و غیر متنی تا چه اندازه قادر به بازسازی منطقه مورد نظر بوده‌اند و چه مقدار خطا در این روند داشته‌اند. همچنین در صورتی که ناحیه ورودی شامل بخش‌های تکرار شونده، مانند نرده یا آجر دیوار، باشند با توجه به خاصیت هیستوگرامی این ویژگی‌ها و نحوه محاسبه آن‌ها، این ویژگی قادر به تشخیص مناسب این مناطق و حذف آن‌ها خواهد بود.

۴.۸.۴ تولید مناطق متنی

برای آموزش کلاسه‌بند آبشاری هم به داده‌های متنی و هم به داده‌های غیر متنی نیاز داریم. با توجه به کثرت داده‌های تصویری در شبکه جهانی وب، تولید داده‌های غیر متنی سخت نیست و می‌توان برای آن از هر تصویری که شامل بخش متنی نباشد استفاده کرد. اما برای بخش داده‌های متنی، با توجه به این که تعداد مناطق متنی و داده‌های موجود در این زمینه محدود است، با الهام از [۴۲] یک الگوریتم تولید تصاویر متنی را مورد استفاده قرار خواهیم داد. این الگوریتم شامل گام‌های زیر است:

انتخاب رنگ

دو رنگ C_f و C_b به ترتیب به عنوان رنگ‌های متنی و پس زمینه از تصاویر آموزشی انتخاب می‌شوند.

---

[۲۶] Absolute Bin Recostruction Error

آماده‌سازی پس‌زمینه

یک بخش تصادفی، p، از یکی از تصاویر آموزشی به صورت تصادفی انتخاب می‌شود. در ادامه هر یک از پیکسل‌های انتخاب شده برای پس‌زمینه با استفاده از رابطه زیر ساخته می‌شوند:

$$p_b(i,j) = \beta \times p(i,j) + (1 - \beta) \times C_b \tag{۲۲.۴}$$

که در آن β پارامتر الگوریتم است که به صورت تصادفی از بازه $[0, 0/2]$ انتخاب می‌شود. دلیل انتخاب این بازه این بوده است که به صورت تجربی هنگامی که اعدادی بزرگتر از . انتخاب می‌شوند پس‌زمینه تصویر بسیار شلوغ شده و مناطق متنی مناسبی به دست نمی‌آیند. چنین ترکیبی باعث می‌شود تا پس‌زمینه تصویر متنی تولید شده به پس‌زمینه‌های متون صحنه‌های طبیعی شباهت پیدا کرده و بخشی از خواص آن‌ها مانند الگوی نویزی آن‌ها را نیز به همراه داشته باشد.

نوشتن متن روی پس‌زمینه

حداکثر ۳ کلمه از یک مجموعه کلمات انتخاب شده و پس از کنار هم قرار داده شدن روی پس‌زمینه تولید شده نوشته می‌شوند. برای تولید تصاویر اعداد نیز یک عدد به صورت تصادفی تولید شده و روی تصویر نوشته می‌شوند.

اضافه کردن نویز گوسی

در این مرحله یک تابع دو بعدی گوسی با پارامترهای تصادفی تولید شده و به تصویر اضافه می‌شود. این مرحله سعی بر شبیه‌سازی اثر فلش روی تصویر دارد. این بخش از الگوریتم فقط برای کم‌تر از ۵ درصد تصاویر تولید شده انجام می‌شود. همچنین برای ۵ درصد از تصاویر، فیلترهای حرکت خطی[۲۷]، محوسازی[۲۸] اعمال می‌شوند. شکل ۹.۴برخی از تصاویر

[۲۷] Linear Motion
[۲۸] Blurring

(آ)

(ب)

شکل ۹.۴: نمونه‌هایی از تصاویر متنی تصنعی تولید شده آ) تصاویر تصنعی انگلیسی ب) تصاویر تصنعی انگلیسی

تولید شده برای متون فارسی و انگلیسی را نشان می‌دهد.

۹.۴ حذف مناطق شبه متنی و ادغام مناطق هم‌پوشان

در آخرین گام الگوریتم تمامی مناطق متنی تشخیص داده شده در سطوح مختلف هرم تصویر در یک مجموعه قرار می‌گیرند. گاهی اوقات برخی مناطق که در آن‌ها برخی خطوط متنی با هم ادغام شده‌اند توسط کلاس‌بند آبشاری به عنوان متن شناخته می‌شوند. ما عبارت مناطق «شبه متنی» را برای این مناطق به کار می‌بریم. برای شناسایی چنین مناطقی، ابتدا ارتفاع اجرای متصل لبه‌ها با ارتفاع کل ناحیه مقایسه می‌شوند؛ در ادامه اگر ارتفاع جزء متصلی کمتر از ۰/۶ ارتفاع کل ناحیه باشد آن جز متصل کنار گذاشته می‌شود.

در نهایت اگر مساحت پوسته محدب[۲۹] مناطق باقیمانده کمتر از نصف مساحت کل ناحیه باشد، ناحیه مورد نظر به عنوان شبه متن در نظر گرفته شده و حذف می‌شود. در نهایت نواحی با هم‌پوشانی بیش از ۳۰ درصد که نسبت ارتفاع آن‌ها بیش از ۵۰ درصد باشد با هم ادغام می‌شوند.

۱۰.۴ نتایج تجربی

برای بررسی کارایی الگوریتم پیشنهادی دو مجموعه داده متفاوت برای تصاویر انگلیسی و فارسی مورد استفاده قرار می‌گیرد. برای زبان انگلیسی مجموعه تصاویر ICDAR 2013 مورد استفاده قرار گرفتند. این مجموعه تصاویر شامل ۴۶۲ تصویر است که از میان آن‌ها ۲۳۳ تصویر مربوط به داده‌های آموزشی و مابقی تصاویر آزمون هستند. تعدادی از تصاویر موجود در این مجموعه داده در شکل ۱۰.۴ نشان داده شده‌اند.

برای فارسی تصاویری از شهرهای مختلف ایران از جمله شاهرود، تبریز، تهران، مشهد، شیراز، اصفهان تهیه شدند. همچنین تصاویری از صحنه‌های مختلف دیگر شامل جعبه‌های اجناس یا کتاب‌های یا سوژه‌های دیگر تهیه شدند. این تصاویر با استفاده از دستگاه‌های مختلف تلفن همراه و حتی دوربین‌های حرفه‌ای تهیه شده‌اند. این تصاویر شامل متون با رنگ‌ها و اندازه‌های متفاوت و ساختارهای متنوع در شرایط نوری مختلف هستند؛ پس زمینه این تصاویر نیز شامل پس‌زمینه‌های یکنواخت و همچنین پس‌زمینه‌های پیچیده هستند. ما این مجموعه داده را فارست[۳۰] می‌نامیم. نمونه‌هایی از تصاویر موجود در این مجموعه داده نیز در شکل ۱۱.۴ نشان داده شده است.

برای ارزیابی روش پیشنهادی روی مجموعه داده $ICDAR$ در ابتدا نسبت هم‌پوشانی میان پنجره پیش‌بینی شده و حقیقت زمینه[۳۱] به دست می‌آید. این مقدار پس از افقی ساختن پنجره‌ها صورت می‌گیرد. برای هر جفت از پنجره‌های پیش‌بینی شده و حقیقت زمینه، یک

[۲۹] Convex Hull
[۳۰] Farset
[۳۱] Ground Truth

(آ)

(ب)

(ج)

(د)

(ه)

(و)

(ز)

(ح)

(ط)

شکل ۴.۱۰: نمونه‌هایی از تصاویر موجود در مجموعه داده ICDAR 2013

(آ) (ب) (ج)

(د) (ه) (و)

(ز) (ح) (ط)

شکل ۱۱.۴: نمونه‌هایی از تصاویر موجود در مجموعه داده فارست

پیش‌بینی در محاسبه نرخ بازشناسی [۳۲] در صورتی که نرخ هم‌پوشانی بیش از ۸۰ درصد باشد صحیح تلقی می‌شود. این مورد برای معیار صحت[۳۳]، ۴۰ درصد در نظر گرفته شده است [۱۰۴، ۱۰۵]. برای ارزیابی روی مجموعه تصاویر فارست نیز روش قانعی و همکاران [۷۷] مورد استفاده قرار گرفت. در این روش ارزیابی در صورتی که نرخ هم‌پوشانی بیش از ۶۰ درصد بوده و زاویه میان دو پنجره کمتر از ۳۰ درصد باشد، تشخیص درست به حساب می‌آید. در بقیه موارد این دو روش همانند هم عمل می‌کنند و معیارهای صحت، بازشناسی و $f - measure$ به صورت زیر تعریف می‌شوند:

$$Precision = \frac{|TP|}{|E|} \tag{۲۳.۴}$$

$$Recall = \frac{|TP|}{|T|} \tag{۲۴.۴}$$

$$F - measure = ۲ * \frac{Precision \times Recall}{Precision + Recall} \tag{۲۵.۴}$$

که در آن $|TP|$ تعداد مناطقی صحیح تشخیص داده شده، $|E|$ تعداد نواحی که توسط الگوریتم متن تشخیص داده شده‌اند و $|T|$ نیز تعداد نواحی حقیقت زمینه را نشان می‌دهد. برای پیاده‌سازی این روش نیز از محیط نرم‌افزار متلب مورد استفاده قرار گرفت. همانگونه که در بخش قبلی نیز مورد اشاره قرار گرفت برای آموزش کلاسه‌بند آبشاری هم تصاویر متنی بریده شده طبیعی و هم تصاویر متنی تصنعی مورد استفاده قرار گرفتند.

برای تولید تصاویر تصنعی دو مجموعه واژگان مورد استفاده قرار گرفتند. برای فارسی مجموعه بیژن‌خان [۱۰۶] مورد استفاده قرار گرفت. این مجموعه شامل مجموعه‌ای از

[۳۲] Recall

[۳۳] Precision

جدول ۱.۴: تأثیر اضافه کردن مجموعه تصاویر متنی تصنعی بر کارایی سیستم پیشنهادی

مجموعه داده	صحت	بازشناسی	f-measure
بریده شده	۸۰/۳۱	۷۸/۳۰	۷۹/۲۹
تصنعی	۹۹/۸۹	۹۲/۳۰	۹۵/۹۵
بریده شده و تصنعی	۹۹/۸۹	۹۴/۴۰	۹۷/۰۷

کلمات نشانه‌گذاری[۳۴] فارسی است که برای پژوهش‌های مرتبط با پردازش زبان طبیعی در زبان فارسی مناسب است. این مجموعه از اخبار روزانه و متون عادی تهیه شده است و شامل مطالبی با موضوعات سیاسی، فرهنگی و مباحث دیگر است. این مجموعه شامل ۲/۶ میلیون کلمه فارسی است که توسط گروه پژوهشی پایگاه داده دانشگاه تهران تهیه شده است. برای زبان انگلیسی نیز یک مجموعه واژگان ۳۰۰ هزارتایی از کلمات به اشتراک گذاشته شده مورد استفاده قرار گرفت.

برای نشان دادن تأثیر اضافه کردن نمونه‌های متنی تصنعی، سه کلاسه‌بند شبکه عصبی MLP با ساختار و ویژگی‌های یکسان اما با مجموعه‌های آموزشی متفاوت آموزش داده شدند. هر یک از این شبکه‌های عصبی دارای یک لایه پنهان با ۳۰ نرون و ۷۶۸ ورودی (تعداد ویژگی‌های لایه سوم کلاسه‌بند آبشاری) بودند. شبکه عصبی اول تنها با استفاده از ۱۰۰۰ تصویر بریده شده متنی آموزش داده شد. شبکه عصبی دوم با استفاده از ۱۰۰ هزار نمونه متنی تصنعی آموزش داده شد و در نهایت شبکه عصبی سوم با مجموعه تصاویر متنی بریده شده و تصنعی آموزش داده شد. برای هر سه کلاسه‌بند تعداد تصاویر غیر متنی، مساوی با تصاویر متنی در نظر گرفته شدند.

کارایی این کلاسه‌بندها با استفاده از یک مجموعه شامل ۱۰۰۰ تصویر متنی و ۱۰۰۰ تصویر غیرمتنی بریده شده مورد سنجش قرار گرفت. تصاویر این مجموعه داده در طی روند آموزش مورد استفاده قرار نگرفته بودند. نتایج این مقایسه در جدول ۱.۴ نشان داده شده است. همانگونه که مشاهده می‌شود استفاده از نمونه‌های آموزشی تصنعی باعث بهبود قابل توجه در نتایج به دست آمده شده است.

نتایج حاصل از پیاده‌سازی روش پیشنهادی روی مجموعه داده $ICDAR2013$ در جدول ۲.۴

[۳۴] Tagged

(آ)

(ب)

(ج)

(د)

(ه)

(و)

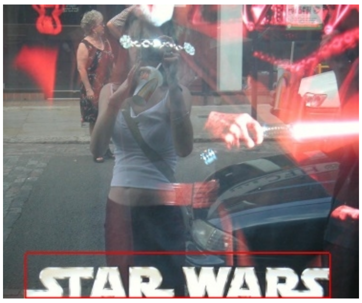
(ز)

(ح)

شکل ۱۲.۴: برخی از خروجی‌های روش پیشنهادی برای مجموعه داده فارست

(آ)

(ب)



(ج)

(د)

(ه)

(و)

(ز)

(ح)

شکل ۱۳.۴: برخی از خروجی‌های روش پیشنهادی برای مجموعه داده فارست

جدول ۲.۴: نتایج روش پیشنهادی روی مجموعه داده ICDAR 2013

صحت	بازشناسی	F-measure
۹۰/۱۰	۷۵/۴۶	۸۲/۱۴

جدول ۳.۴: نتایج روش پیشنهادی روی مجموعه تصاویر فارست

صحت	بازشناسی	F-measure
۹۰/۸۴	۸۷/۵۴	۸۹/۱۶

نشان داده شده است. همانگونه که مشاهده می‌شود روش پیشنهادی نتایج خوبی روی این مجموعه داده تولید کرده است. همچنین نتایج پیاده‌سازی روش پیشنهادی روی مجموعه داده فارست در جدول ۳.۴ آمده است. نتایج نشان داده شده در جدول ۲.۴ و جدول ۳.۴ بیانگر این نکته هستند که روش پیشنهادی هم برای مجموعه تصاویر فارسی و هم مجموعه تصاویر انگلیسی کارایی مناسبی از خود نشان داده است. تعدادی از خروجی‌های تولید شده توسط روش پیشنهادی نیز در شکل ۱۲.۴ و شکل ۱۳.۴ به نمایش در آمده‌اند.

۱۱.۴ جمع‌بندی

در این فصل «امضای رنگ لبه‌ها» به عنوان اولین روش ارائه شده در این فصل مورد بررسی قرار گرفت. این روش با استخراج لبه‌های تصویر ورودی کار را شروع کرده و در ادامه برای هر یک از پیکسل‌های لبه، رنگ‌های اطراف آن را به کمک الگوریتم انتقال میانگین به دست می‌آورد که آن‌ها را امضای رنگ لبه‌ها می‌نامیم؛ در مرحله بعد با استفاده از امضای رنگ لبه‌ها و الگوریتم خوشه‌بندی اتصال‌ـتکی پیکسل‌های با رنگ مشابه با هم گروه‌هایی را تشکیل می‌دهند که هر یک از این گروه‌ها به عنوان یک ناحیه کاندیدای متنی در نظر گرفته می‌شوند. این نواحی با استفاده از یک کلاسه‌بند آبشاری با چند ویژگی نو کلاسه‌بندی می‌شوند. این روش دقت مناسبی دارد امّا با توجه به آن که در آن از دو الگوریتم خوشه‌بندی متفاوت استفاده شده است به لحاظ محاسباتی هزینه بالایی دارد که در فصل بعدی به این مشکلات پرداخته و آن‌ها را رفع می‌کنیم.

فصل ۵

تبدیل رنگ لبه‌ها

۱.۵ مقدمه

مفهوم مورد استفاده و روند کلی الگوریتم مورد استفاده در این فصل همانند فصل ۴ است؛ امّا الگوریتم‌های مورد استفاده در بخش‌های مختلف آن متفاوت بوده و بهبود یافته‌اند. موارد بهبود روش جاری نسبت به روش ارائه شده در فصل ۴ به شرح زیر است:

- در این روش الگوریتم استخراج لبه مورد استفاده بهبود داده می‌شود.

- فرآیند به دست آوردن رنگ‌های مربوط به لبه‌ها با استفاده از روشی به نام «تبدیل رنگ لبه»[1] بهبود داده می‌شود. این روش هم روی صحت رنگ‌های به دست آمده از لبه تأثیرگذار است و هم بسیار سریع‌تر از الگوریتم انتقال میانگین عمل می‌کند.

- روش اتصال لبه‌های تصویر نیز در این روش بهبود داده شده است. به لحاظ مفهومی

[1] Edge Color Transform

روش ارائه شده در این فصل مشابه با الگوریتم خوشه‌بندی اتصال‌ـ تکی عمل می‌کند؛ امّا از لحاظ سرعت بسیار بهتر از آن است.

۲.۵ به دست آوردن لبه ها

همانگونه که در فصل ۴ اشاره شد، برای به دست آوردن لبه مناطق متنی روش سوبل بهتر از کنی عمل می‌کند؛ امّا با توجه به اینکه روش معمول برای به دست آوردن لبه‌های یک تصویر تبدیل آن به یک تصویر خاکستری[۲] و سپس استخراج لبه‌هاست، گاهی در این روند اطلاعات با ارزش رنگی که می‌توان از طریق آن‌ها لبه‌ها را استخراج کرد از بین می‌روند. در واقع گاهی در تصویر رنگی اولیه، کنتراست خوبی میان متن و پس‌زمینه آن وجود دارد که پس از تبدیل یک تصویر رنگی به خاکستری این کنتراست تضعیف می‌شوند و لبه‌ها به خوبی استخراج نمی‌شوند. در حالت کلی ممکن است فقط در یک کانال رنگی کنتراست خوبی بین متن و پس‌زمینه وجود داشته باشد و با ترکیب شدن سه کانال رنگی این کنتراست شدت اولیه خود را از دست بدهد.

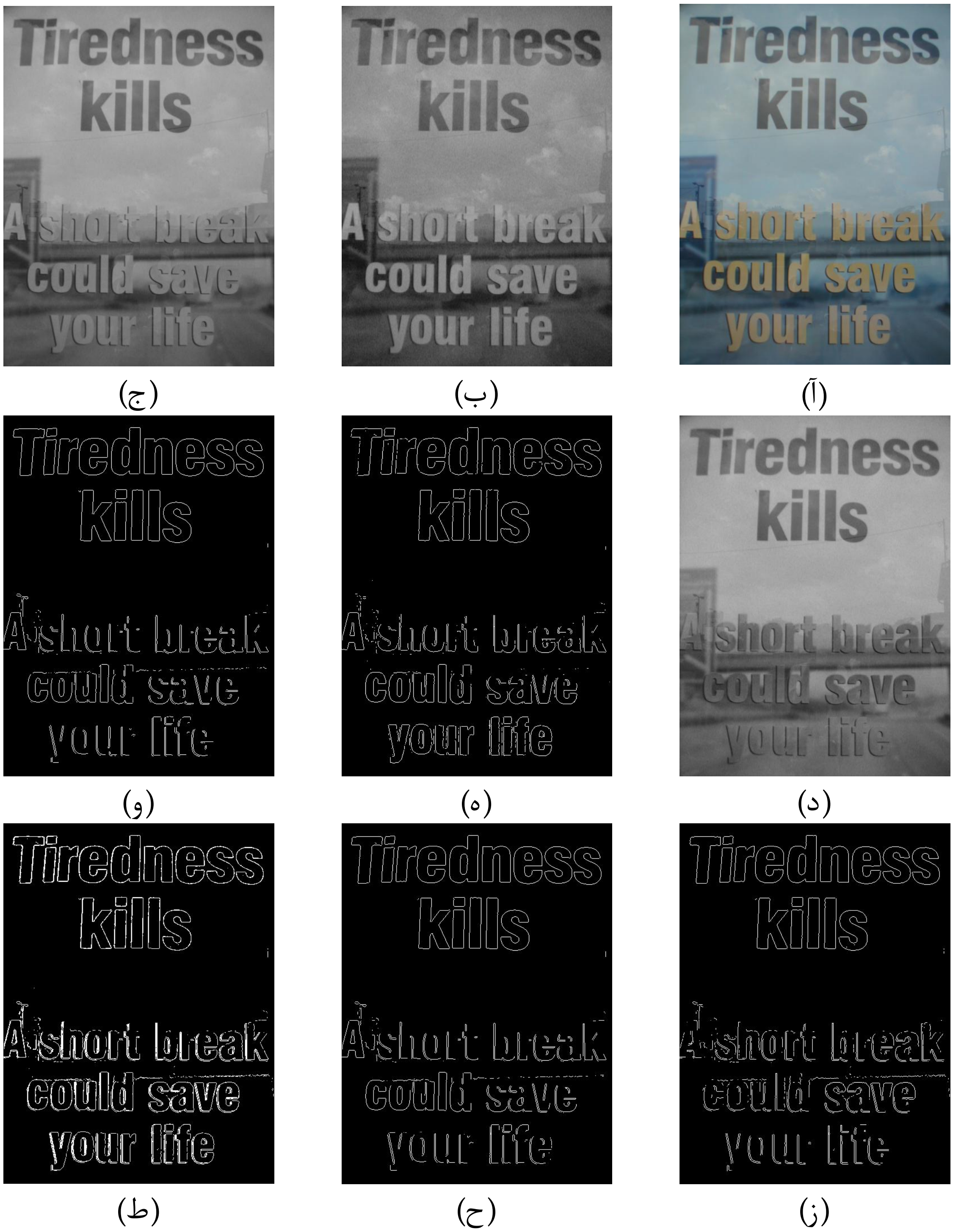
این مورد در شکل ۱.۵ نشان داده شده است. همانگونه که در این شکل نیز مشاهده می‌شود در هیچ یک از کانال‌های رنگی لبه‌های بخش متنی به صورت کامل و درست استخراج نشده‌اند و هر یک در بخش‌هایی موفق‌تر از دو کانال دیگر بوده‌اند؛ اما لبه‌های حاصل از ترکیب این سه کانال رنگی بسیار بهتر از تک تک آن‌ها است.

۳.۵ تبدیل رنگ لبه‌ها

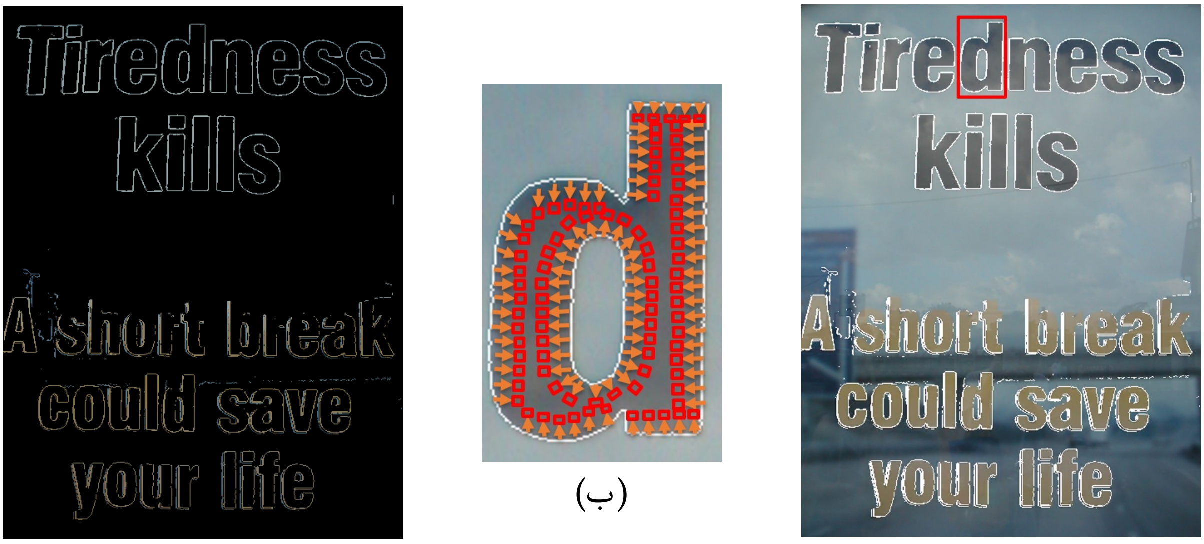
پس از یافتن لبه‌های تصویر ورودی همانند روش ارائه شده در فصل ۴ در اینجا نیز روشی برای یافتن امضای رنگ لبه‌ها ارائه می‌کنیم. روش ارائه شده در این بخش دو مزیت نسبت به الگوریتم خوشه‌بندی انتقال‌ـ میانگین دارد:

- این روش بسیار سریع‌تر از روش انتقال میانگین عمل می‌کند و نیازی به استفاده از

[۲] Gray Scale



شکل ۱.۵: تاثیر ترکیب خروجی لبه‌های به دست آمده از سه کانال رنگی قرمز، سبز و آبی روی کیفیت لبه‌ها آ) تصویر ورودی ب) کانال رنگی قرمز ج) کانال رنگی سبز د) کانال رنگی آبی ه) لبه‌های کانال قرمز و) لبه‌های کانال سبز ز) لبه‌های کانال آبی ح) لبه‌های تصویر خاکستری ط) لبه‌های حاصل از ترکیب سه کانال رنگی



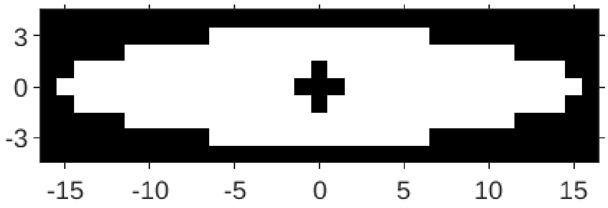
(ج) (ب) (آ)

شکل ۲.۵: نحوه عملکرد تبدیل رنگ لبه‌ها آ) تصویر ورودی که در آن لبه‌ها با رنگ سفید مشخص شده‌اند ب) نحوه به دست آوردن امضای رنگ لبه‌ها ج)یکی از خروجی‌های عملگر تبدیل رنگ لبه‌ها

یک همسایگی $n \times n$ برای به دست آوردن رنگ‌های اطراف ناحیه متنی ندارد.

- این روش تنها دو رنگ برای هر لبه استخراج می‌کند که یکی از آن‌ها برای پس‌زمینه و دیگری برای پیش‌زمینه است؛ داشتن تنها دو رنگ به ساده‌سازی مراحل بعدی الگوریتم کمک شایانی می‌کند و نیاز به استفاده از عملگر کمینه را برای به دست آوردن فاصله رنگی دو پیکسل از بین می‌برد.

در این روش برای هر پیکسل لبه p به اندازه m پیکسل در جهت گرادیان حرکت می‌کنیم و به یک پیکسل جدید q می‌رسیم. در ادامه، رنگ پیکسل q در تصویر اصلی به پیکسل لبه انتساب داده می‌شود. اگر پیکسل لبه در مرز میان محدوده متنی و پس‌زمینه قرار داشته باشد رنگ انتخاب شده برای آن، بسته به این که متن از پس‌زمینه روشن‌تر است یا تاریک‌تر، یا رنگ متن خواهد بود و یا رنگ پس‌زمینه‌ای که متن روی آن قرار دارد. تمامی این روند در جهت عکس گرادیان نیز تکرار می‌شود؛ بنابراین حاصل این روند دو تصویر لبه رنگی‌خواهد بود. خروجی این روند برای یک تصویر نمونه در شکل ۲.۵ نشان داده شده است.



شکل ۳.۵: یک همسایگی با مقدار ۱۵ = δ که مرکز همسایگی با علامت + نشان داده شده است

۴.۵ الگوریتم رشد لبه

پس از اعمال الگوریتم «تبدیل رنگ لبه‌ها» و به دست آوردن دو تصویر لبه رنگی، الگوریتمی به نام «رشد لبه»[۳] برای ایجاد نواحی کاندیدای متنی پیشنهاد می‌گردد. در این الگوریتم نقاط لبه‌ای که در یک همسایگی از پیش تعریف شده از همدیگر قرار داشته باشند به عنوان همسایه در نظر گرفته می‌شوند. به عبارت دیگر پیکسل‌های r و s اگر شرایط زیر را ارضا کنند به عنوان همسایه در نظر گرفته می‌شوند:

$$|x_r - x_s| + |y_r - y_s|^{۲} \leq \delta \quad (۱.۵)$$

که در آن x و y مختصات یک پیکسل در تصویر ورودی را نشان می‌دهند و δ پارامتر الگوریتم است که اندازه همسایگی را کنترل می‌کند. مقادیر بزرگتر δ منجر به همسایگی‌های بزرگتری می‌شوند. در طراحی این همسایگی دو نکته مورد توجه قرار گرفته‌اند: اول اینکه معمولاً کلمات موجود در متون در کنار همدیگر قرار دارند و دوم اینکه معمولاً در بالا یا پایین متون موجود در تصاویر نیز متن‌های دیگری قرار دارند؛ بنابراین ناحیه همسایگی مورد نظر باید به گونه‌ای طراحی شود که هم بتواند کلمات موجود در یک خط را به هم وصل کند و هم بتواند تا حد امکان از الحاق خطوطی متنی به همدیگر جلوگیری کند. با در نظر گرفتن این موارد ناحیه متنی به صورت بیضی‌وار طراحی گردید و توان دوم مورد استفاده در نامساوی (۱.۵) به همین منظور است. شکل ۳.۵ نمونه‌ای از نقاط داخل این همسایگی را برای ۱۵ = δ نشان می‌دهد.

[۳] Edge Growing

همچنین فاصله میان دو پیکسل را به صورت زیر تعریف می‌کنیم:

$$d(p, q) = \sqrt{(r'_p - r'_q)^۲ + (g'_p - g'_q)^۲ + (b'_p - b'_q)^۲} \quad (۲.۵)$$

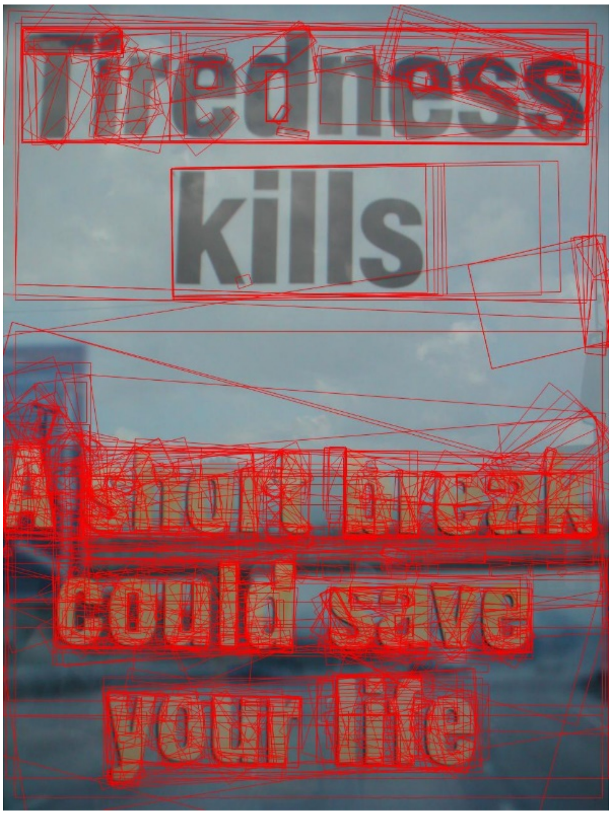
$$r' = \frac{r}{r + g + b} \quad (۳.۵)$$

$$g' = \frac{g}{r + g + b} \quad (۴.۵)$$

$$b' = \frac{b}{r + g + b} \quad (۵.۵)$$

که در آن r، g و b به ترتیب نشان دهنده کانال‌های قرمز، سبز و آبی پیکسل‌ها هستند. نرمال‌سازی مقادیر رنگی هر یک پیکسل‌های لبه باعث می‌شوند تا الگوریتم در برابر تغییرات شدت روشنایی مقاومت بیشتری از خود نشان دهد. الگوریتم رشد لبه با استفاده از یک صف پیاده‌سازی می‌شود؛ در ابتدا یکی از نقاط لبه به صف اضافه می‌شود. پس از آن مراحل بعدی تکرار می‌شوند. یک پیکسل از صف برداشته شده و همسایه‌های آن با استفاده از فرمول (۱.۵) به دست می‌آیند. برای هر یک از همسایه‌ها فاصله رنگی میان آن و میانگین رنگ پیکسل‌هایی که تاکنون در تکرار جاری به صف اضافه شده محاسبه می‌شود و در صورتی که از حد مشخصی پایین‌تر باشد آن پیکسل به صف اضافه می‌شود؛ این روند تا زمانی ادامه پیدا می‌کند که صف خالی شود. تمامی پیکسل‌هایی که به صف اضافه شده و از آن برداشته شده‌اند در نهایت یک ناحیه کاندیدای متنی را نشان می‌دهند. خروجی این روش برای یک تصویر نمونه در شکل ۴.۵ نشان داده شده است.

پس از خالی شدن صف، از میان پیکسل‌هایی که تاکنون ملاقات نشده‌اند یک پیکسل انتخاب شده و به صف اضافه می‌شود و روند فوق برای پیکسل جدید نیز تکرار می‌شود تا تمامی



شکل ۴.۵: نواحی استخراج شده توسط الگوریتم رشد لبه برای یک تصویر نمونه

پیکسل‌های لبه موجود در تصویر ملاقات شوند. این الگوریتم در شکل ۵.۵ نشان داده شده است.

معمولاً مرتبه زمانی الگوریتم‌هایی که برای پردازش تصویر به کار می‌روند با توجه به عرض و ارتفاع تصویر گزارش می‌شود؛ امّا با توجه به این که الگوریتم رشد لبه و همچنین الگوریتم خوشه‌بندی اتصال‌ـ‌تکی که در قالب روش امضای رنگ لبه‌ها مورد استفاد قرار گرفت هر دو فقط از پیکسل‌های لبه استفاده می‌کنند و تعداد این پیکسل‌ها برای تصاویر مختلف، متفاوت است، تحلیل الگوریتم‌ها را با توجه به تعداد پیکسل‌های لبه آن‌ها انجام می‌دهیم. اگر تعداد پیکسل‌های لبه به دست آمده از یک تصویر ورودی را n در نظر بگیریم. در این صورت می‌توان با بررسی تحلیلی الگوریتم نشان داده شده در شکل ۵.۵ مرتبه زمانی آن را به دست آورد. خطوط ۵ تا ۹ الگوریتم همگی $O(1)$ هستند. خطوط ۱۲ تا ۱۷ الگوریتم نیز همگی حداکثر به تعداد پیکسل‌های موجود در همسایگی تکرار می‌شوند که تعدادی ثابت بوده و در نتیجه $O(1)$ هستند. برای بررسی برچسب‌دار بودن یا نبودن پیکسل‌ها نیز

Input: ECTMap: output of ECT operator
γ: threshold for merging pixels in the same neighborhood

Output: labelMap: a map which assigns a label number to each pixel of ECTMap

Functions:
remove(a): removes an item from list, a, and returns it
rgb(a): returns the rgb values of pixel a
position(a): returns spatial position of pixel a

Body of Algorithm:
1. $pixelList = \{set\ of\ all\ of\ nonzero\ pixels\ in\ ECTMap\}$;
2. $lableCounter = 1$;
3. $q = \{\}$;
4. **While** pixelList is not empty do
5. $p = remove\ (pixelList)$;
6. $q = q \cup \{p\}$;
7. $avgColor = rgb(p)$;
8. $n = 1$;
9. $labelMap(position(p)) = labelCounter$;
10.   **While** q is not empty do
11. $r = remove\ (q)$;
12.   **for each** $s \in neigboorhood(r)$ and $s \in pixelList$ do
13.   **if** $\|rgb(s) - avgColor\|_2 \leq \gamma$
14. $labelMap(position(s)) = labelCounter$;
15. $q = q \cup \{s\}$;
16. $avgColor = \frac{n \times avgColor + rgb(s)}{n+1}$;
17. $n = n + 1$;

شکل ۵.۵: الگوریتم رشد لبه

می‌توان از یک نقشه به اندازه تصویر ورودی استفاده کرد و در زمان ثابت بررسی کرد که آیا آن پیکسل قبلاً برچسب خورده یا نه. بنابراین تمامی دستورات در حلقه سطر ۴ از مرتبه زمانی $O(1)$ هستند و از آن‌جا که این حلقه حداکثر به تعداد پیکسل‌های لبه تکرار می‌شوند می‌توان نتیجه گرفت که این الگوریتم از مرتبه زمانی $O(n)$ است. در طرف مقابل می‌توان نشان داد که الگوریتم خوشه‌بندی اتصال‌ـ‌تکی در بهترین حالت از مرتبه زمانی $O(n^2)$ است [۱۰۷]. بنابراین به لحاظ مرتبه زمانی الگوریتم رشد لبه مزیت قابل توجهی نسبت به روش خوشه‌بندی اتصال‌ـ‌تکی خواهد داشت.

۵.۵ نتایج به دست آمده

کلاسه‌بندی آبشاری فصل ۴ در این فصل نیز برای کلاسه‌بندی مورد استفاده قرار گرفت. نتایج به دست آمده از پیاده‌سازی این روش روی مجموعه داده‌های $ICDAR2013$ و فارست به ترتیب در جدول‌های ۱.۵ و ۲.۵ نشان داده شده است. همچنین تعدادی از خروجی‌های روش پیشنهادی دوم در شکل ۶.۵ برای تصاویر انگلیسی و شکل ۷.۵ فارسی نشان داده شده است.

جدول ۱.۵: نتایج روش پیشنهادی روی مجموعه داده ICDAR 2013

صحت	بازشناسی	F-measure
۹۳/۸۵	۷۷/۷۷	۸۵/۰۲

جدول ۲.۵: نتایج روش پیشنهادی روی مجموعه تصاویر فارست

صحت	بازشناسی	F-measure
۹۲/۰۲	۸۸/۱۶	۹۰/۰۵

۶.۵ جمع‌بندی

در این فصل «تبدیل رنگ لبه‌ها» به عنوان دومین روش پیشنهادی ارائه شد. این روش با ارائه یک عملگر جدید رنگ‌های اطراف لبه‌ها را به صورت دقیق و بسیار سریع استخراج می‌کند. در ادامه الگوریتم جدید دیگری با نام «رشد لبه» معرفی گردید که می‌تواند بسیار سریع‌تر از الگوریتم خوشه‌بندی اتصال‌ـتکی به گروه‌بندی نقاط لبه و ایجاد نواحی کاندیدای متنی بپردازد. این روش از همان کلاسه‌بند آبشاری روش امضای رنگ لبه‌ها استفاده می‌کند و با توجه به تغییرات ایجاد شده در روش تولید نواحی کاندیدای متنی نتایج بهتری تولید می‌کند.

شکل ۶.۵: برخی از خروجی‌های روش پیشنهادی برای مجموعه داده $ICDAR2013$

(آ)

(ب)

(ج)

(د)

(ه)

(و)

(ز)

(ح)

شکل ۷.۵: برخی از خروجی‌های روش پیشنهادی برای مجموعه داده فارست

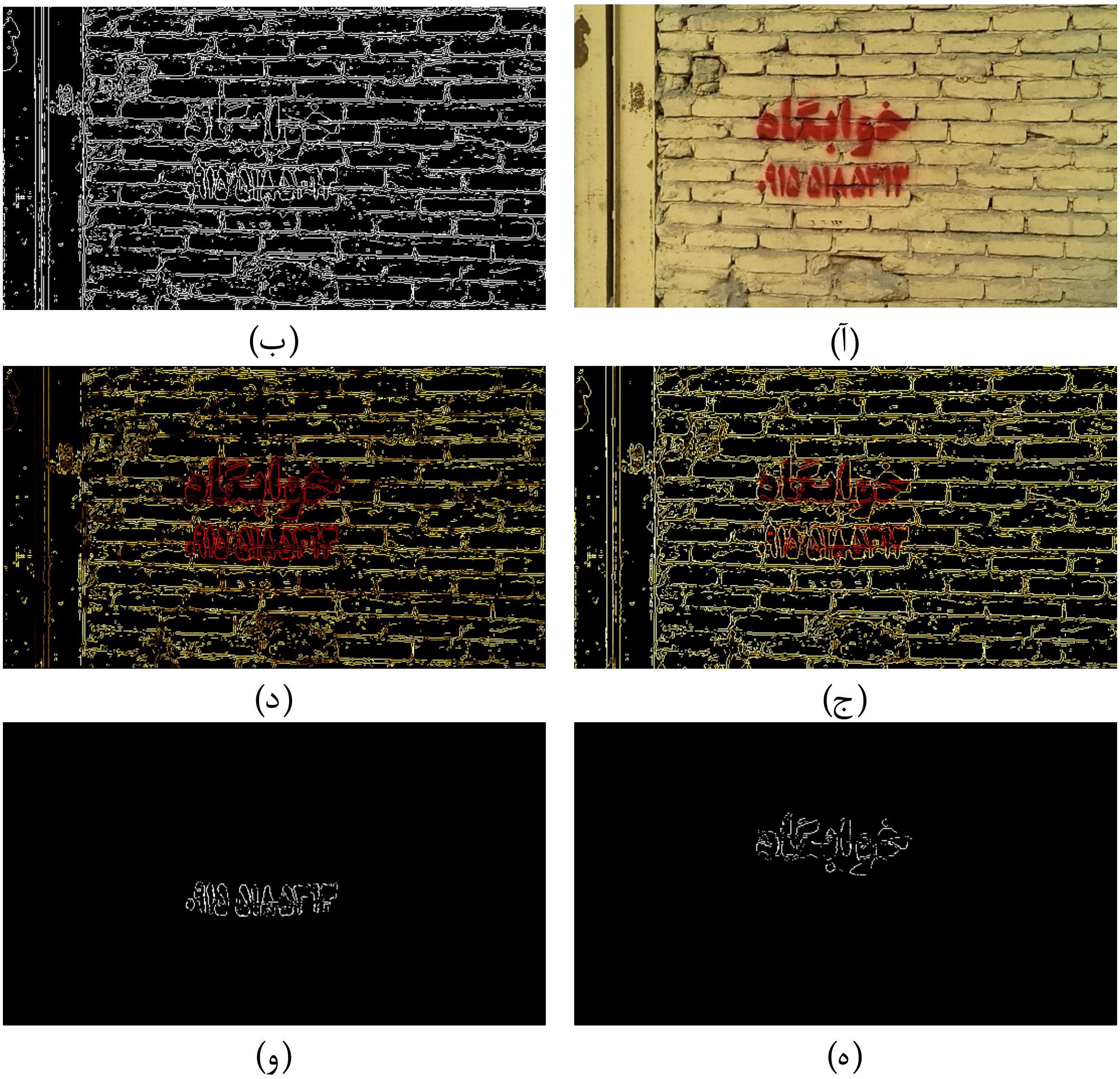
فصل ۶

تبدیل رنگ عمیق و شبکه فرا اکتشافی

۱.۶ مقدمه

ترکیب «تبدیل رنگ لبه» به همراه الگوریتم «رشد لبه» در بسیاری موارد از جمله مواردی که پس‌زمینه متن پیچیده باشد کارایی خوبی از خود نشان می‌دهد. به عنوان مثال خروجی این الگوریتم برای یک تصویر نمونه در شکل ۱.۶ نشان داده شده است. همانگونه که در این شکل نیز دیده می‌شود خروجی به دست آمده برای این تصویر با پس‌زمینه پیچیده بسیار خوب بوده است؛ با دقت در شکل ۱.۶ب مشاهده می‌شود که به سختی می‌توان لبه‌های بخش متنی و غیرمتنی را از هم تمیز داد؛ امّا الگوریتم با استفاده از لبه‌های رنگی توانسته است به خوبی بین این بخش‌ها با استفاده از رنگ آن‌ها تمایز قائل شود.

با وجود اینکه روش ارائه شده در فصل ۵ کارایی خوبی از خود نشان می‌دهد امّا معایبی نیز دارد که در این فصل به آن‌ها اشاره شده و سعی می‌کنیم آن‌ها را مرتفع سازیم؛ همچنین در این فصل به جای کلاسه‌بند آبشاری مورد استفاده در فصل ۴ یک شبکه کانولوشنی با

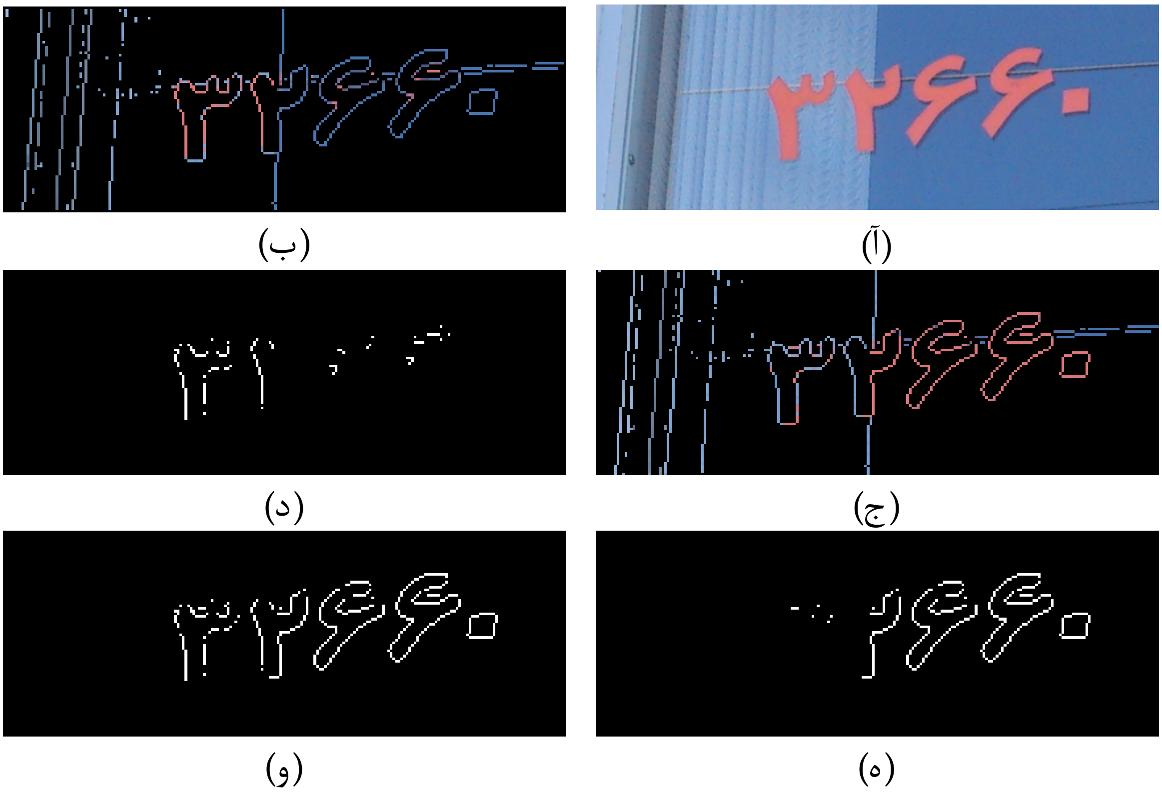


شکل ۱.۶: خروجی تبدیل رنگ لبه‌ها و الگوریتم رشد لبه برای یک تصویر با پس‌زمینه پیچیده (برای بهبود کیفیت، کنتراست تصاویر افزایش داده شده است) آ) تصویر ورودی ب) لبه‌های تصویر ورودی ج_د) دو تصویر لبه رنگی ه_و) خروجی الگوریتم برای بخش‌های متنی

ساختاری جدید معرفی می‌شود.

۲.۶ بهبود الگوریتم فصل گذشته

اولین مشکل روش فصل ۵ در مواردی مانند تصویر نشان داده شده در شکل ۲.۶ نمود پیدا می‌کند. در این تصویر بخشی از پس‌زمینه متن از آن روشن‌تر بوده و بخش دیگر از آن تاریک‌تر است؛ بنابراین جهت گرادیان برای بخشی از پیکسل‌های لبه‌های متنی به سمت داخل متن و برای برخی پیکسل‌ها به سمت خارج متن است؛ این موضوع باعث می‌شود



شکل ۲.۶: اثر پس‌زمینه غیریکنواخت در خروجی الگوریتم رشد لبه آ) برشی از بخش متنی یک تصویر با پس‌زمینه غیر یکنواخت ب–ج) خروجی‌های تبدیل رنگ لبه‌ها د–ه) بخش‌های متنی جدا شده در هر بخش و) نتیجه ترکیب نواحی موجود در تصاویر بخش‌های «د» و «ه»

در تصاویر لبه رنگی ایجاد شده، رنگ لبه‌های متنی یکسان نباشد چرا که با توجه به جهت گرادیان، این رنگ‌ها از داخل و یا خارج بخش متنی گرفته شده‌اند. این روند موجب تقسیم شدن ناحیه متنی به دو بخش شده و به صورت یکجا یک گروه لبه کاندیدا را تشکیل نمی‌دهد؛ در نتیجه بخشی از ناحیه متنی در یک تصویر لبه رنگی و بخشی دیگر در تصویر لبه رنگی دیگر تشکیل می‌شود.

راه‌کار پیشنهادی برای رفع این مشکل آن است که پس از تشکیل نواحی کاندیدا از دو تصویر لبه رنگی، ضمن حفظ نواحی کاندیدای موجود، نواحی جدیدی نیز با ترکیب نواحی کاندیدای همسایه‌ای که رنگ مشابه دارند تشکیل شود. در اینجا همسایگی همانند رابطه (۱.۵) تعریف می‌شود؛ با این تفاوت که به جای محاسبه فاصله دو پیکسل، باید فاصله مجموعه‌ای از پیکسل‌ها محاسبه شود. همچنین در اینجا فاصله میان دو ناحیه کاندیدا، فاصله نزدیکترین پیکسل‌های آن‌ها از هم تعریف می‌شود. این تعریف با همسایگی تعریف شده برای دو پیکسل در رابطه (۱.۵) سازگاری کامل دارد و همان نواحی کاندیدایی را تولید

می‌کند که در صورت هموار بودن پس‌زمینه تصویر به دست می‌آوردیم. خروجی الگوریتم اصلاح شده در شکل ۲.۶و نشان داده شده است. همانگونه که در این شکل نیز دیده می‌شود با اعمال راه‌حل پیشنهادی، ناحیه متنی که تکه تکه شده بود به صورت یکپارچه تشخیص داده شده است.

مسئله دیگری که در مورد الگوریتم ارائه شده در فصل ۵ وجود دارد تولید شدن نواحی کاندیدای تکراری است. این مشکل به دو دلیل ایجاد می‌شود؛ اول آن که الگوریتم پیشنهادی برای آن که بتواند مقیاس‌پذیر باشد هرمی از تصویر ورودی را ایجاد می‌کند، امّا این ممکن است باعث شود که نواحی کاندیدا در چندین لایه هرم به صورت تکراری ایجاد شوند؛ همچنین نواحی کاندیدا از دو تصویر لبه رنگی استخراج می‌شوند که ممکن است باز هم یک ناحیه در هر دو تصویر تشخیص داده شود. برای رفع این مشکل پس از تجمیع نواحی کاندیدای متنی به دست آمده، نواحی تکراری شناسایی شده و حذف می‌شوند تا در ادامه مراحل الگوریتم فقط نواحی غیر تکراری شرکت داده شوند.

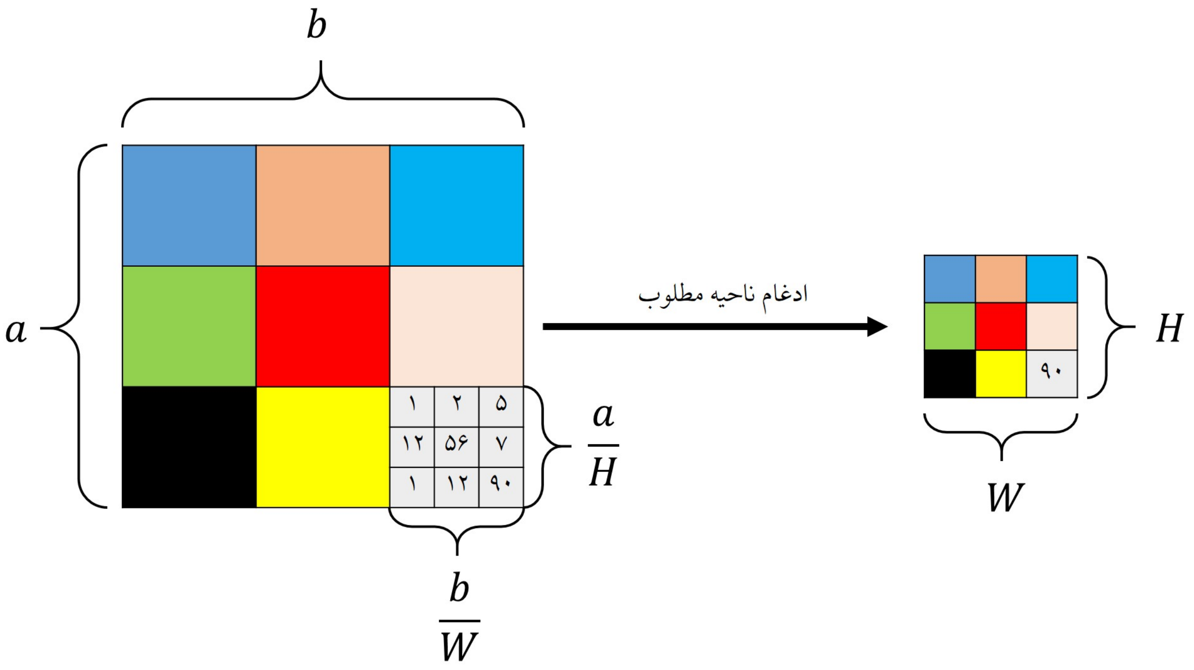
۳.۶ شبکه عصبی کانولوشنی برای تشخیص نواحی متنی

در ساختار شبکه کانولوشنی پیشنهادی از ایده $FastRCNN$ [۲] برای تمایز مناطق متنی از غیرمتنی الهام گرفته شده است. شبکه $FastRCNN$ یک روش مکان‌یابی اشیاء بر اساس شبکه‌های عصبی کانولوشنی است. این روش نیازمند راهکاری برای تولید مناطق کاندیدا است و از دو بخش اصلی تشکیل شده است. بخش اول یک شبکه کانولوشنی است که برای استخراج ویژگی مورد استفاده قرار می‌گیرد و بخش دوم شامل یک لایه برای نگاشت مناطق کاندیدا به فضای ویژگی با طول ثابت و کلاسه‌بندی آن‌ها است.

این شبکه، در ابتدا تصویر ورودی و مجموعه‌ای از مناطق کاندیدا را به عنوان ورودی دریافت می‌کند. پس از اعمال بخش کانولوشنی شبکه روی تصویر ورودی، مجموعه ویژگی‌های کانولوشنی تولید می‌شوند. با توجه به اینکه در روند ایجاد مجموعه ویژگی‌ها، لایه‌های کانولوشن و ادغام بیشینه روی تصویر ورودی اعمال می‌شوند، ابعاد تصویر ورودی پس از

اعمال بخش کانولوشنی شبکه تغییر پیدا می‌کند؛ بنابراین نواحی کاندیدای ورودی نیز باید روی مجموعه ویژگی‌ها نگاشت شوند. این نگاشت شامل چند تقسیم ساده روی مختصات ناحیه کاندیدا است. مثلاً اگر پس از اعمال بخش کانولوشنی یک شبکه ابعاد کانال‌های خروجی نصف شده باشد، مختصات گوشه‌های نواحی کاندیدا نیز نصف می‌شوند.

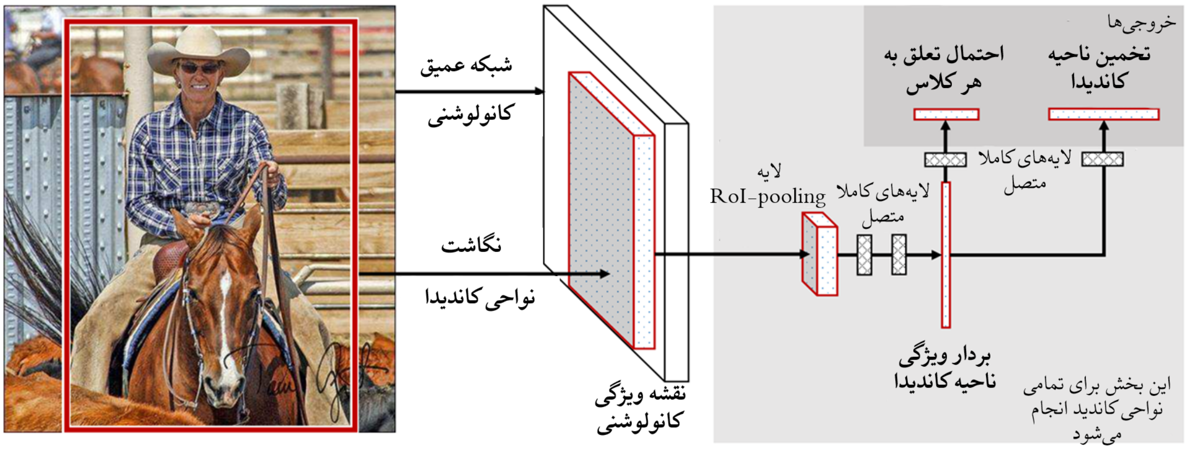
نکته مهمّ دیگری که در مورد این شبکه وجود دارد لایه «ادغام ناحیه مطلوب»[۱] است. از آنجا که ابعاد نواحی کاندیدا با هم تفاوت دارد و لایه‌های بعدی شبکه، لایه‌های کاملاً-متصل بوده و نیاز به ورودی با طول ثابت دارند، از این لایه برای ایجاد مجموعه ویژگی‌های با طول ثابت برای هر یک از نواحی کاندیدا استفاده می‌شود. این لایه، نگاشت ناحیه کاندیدا را در هر یک از کانال‌های ورودی به $H \times W$ بخش تقسیم کرده و در هر بخش مقدار بیشینه را به عنوان یک ویژگی انتخاب می‌کند، که مقادیر H و W پارامترهای این روش هستند. جزئیات نحوه عملکرد این روش در شکل ۳.۶ برای یک کانال نشان داده شده است. روند نشان داده شده در این شکل برای تمامی کانال‌های ورودی لایه اعمال می‌شود؛ به عبارت دیگر این لایه تعداد کانال‌های نگاشت ورودی خود را تغییر نمی‌دهد.



شکل ۳.۶: جزئیات لایه ادغام ناحیه مطلوب برای یک کانال

---
[۱] Region of Interest Pooling

در ادامه، ویژگی‌های به دست آمده از هر ناحیه به یک شبکه کاملاً_ متصل با دو سر خروجی داده می‌شوند. یکی از این دو سر احتمال تعلق هر ناحیه را به هر یک از کلاس‌ها (به همراه یک کلاس پس‌زمینه) تخمین می‌زند و سر دیگر وظیفه بهبود ناحیه کاندیدا را بر عهده داشته و تلاش می‌کند تا ناحیه کاندیدا را به بهترین شکل ممکن بر شیء تشخیص داده شده محاط کند. شمای کلی این روش در شکل ۴.۶ قابل مشاهده است.



شکل ۴.۶: جزئیات شبکه *FastRCNN* [۲]

همانگونه که اشاره شد روش *FastRCNN* نیاز به یک مولّد نواحی کاندیدا دارد؛ در روش پیشنهادی این فصل، از الگوریتم ارائه شده در فصل ۵ به همراه بهبودهای ارائه شده در این فصل به عنوان مولّد نواحی کاندید استفاده می‌شود. همچنین برای جداسازی مناطق متنی و غیرمتنی یک شبکه کانولوشنی با معماری اختصاصی بر پایه *FastRCNN* ارائه می‌شود.

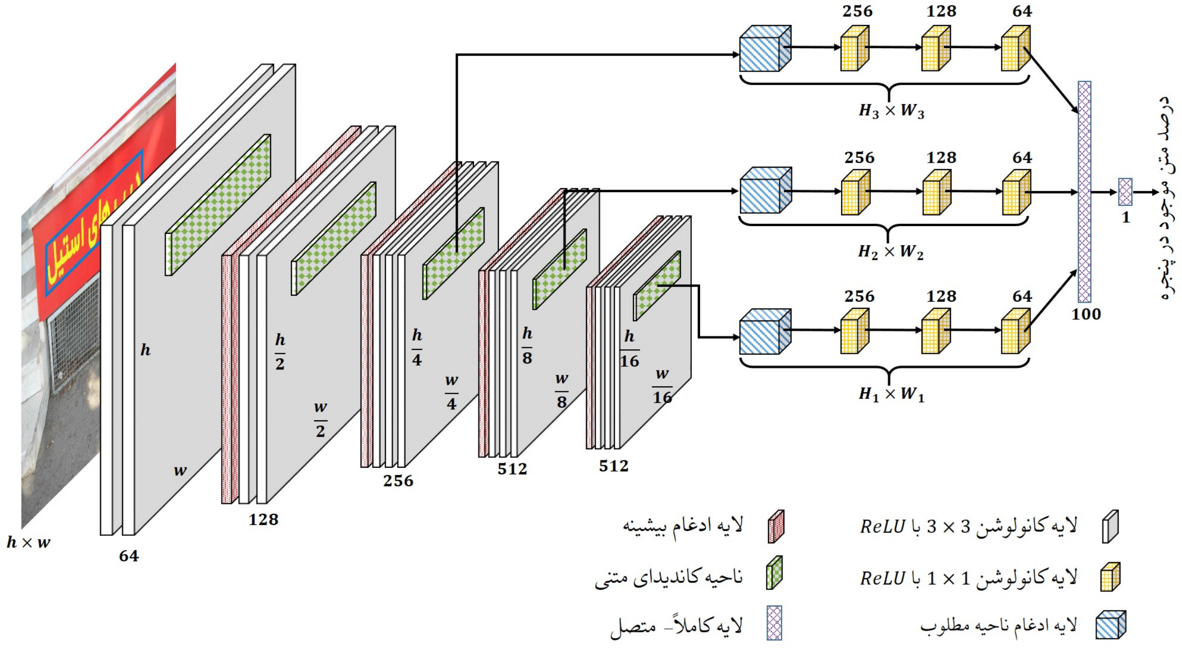
شبکه پیشنهادی در ابتدا همانند *FastRCNN*، پنج مجموعه لایه کانولوشنی *VGG*16 [۲۱] را، که ساختار آن در شکل ۵.۶ نشان داده شده است، روی تصویر ورودی اعمال می‌کند که لایه‌های کاملاً متصل و لایه ادغام بیشینه پس از آخرین لایه کانولوشنی حذف شده‌اند؛ امّا برخلاف *FastRCNN* تنها داری یک سر بوده و شامل سر مربوط به بهبود ناحیه کاندیدا نیست. این تصمیم به این دلیل اتخاذ شده است که مجموعه نواحی کاندیدای تولید شده توسط الگوریتم «تبدیل رنگ لبه‌ها» در بیشتر موارد دارای کیفیت مناسبی بوده و نیازی به بهبود ندارند. حذف این سر از شبکه، کاهش بار محاسباتی شبکه را به دنبال دارد و همچنین تعداد پارامترهای آن را کاهش می‌دهد.

شکل ۵.۶: معماری شبکه $VGG16$

تفاوت دیگر شبکه پیشنهادی با $FastRCNN$ به نحوه استخراج ویژگی از نواحی کاندیدا مربوط می‌شود؛ در روش پیشنهادی لایه ادغام ناحیه مطلوب همزمان روی مجموعه لایه‌های $Conv5$، $Conv4$ و $Conv3$ در ناحیه کاندیدا اعمال می‌شود. دلیل این کار آن است که در لایه آخر شبکه، با توجه به کوچک شدن هر یک از کانال‌های کانولوشنی، جزئیات نواحی کاندیدا از بین می‌رود و نواحی با اندازه‌های متفاوت ممکن است به یک ناحیه کانولوشنی یکسان در لایه آخر نگاشت شوند؛ از طرف دیگر با توجه به اینکه انتظار داریم شبکه برای نواحی کاندیدایی که بهتر نواحی متنی را در برگرفته‌اند، خروجی بزرگتری تولید کند، استفاده از لایه $Conv5$ به تنهایی مشکل ساز می‌شود.

پس از اعمال هر یک از لایه‌های ادغام ناحیه مطلوب، سه لایه کانولوشنی ۱ × ۱ به خروجی

آن‌ها اعمال می‌شود. این لایه‌ها همزمان دارای دو کارکرد هستند: هم تعداد کانال‌های ناحیه کاندیدا را کاهش می‌دهند و هم استفاده از لایه $ReLU$ پس از هر یک از آن‌ها، باعث افزایش قدرت شبکه می‌شود. در ادامه، خروجی حاصل از اعمال لایه‌های کانولوشنی ۱ × ۱ به شکل یک بردار در کنار هم قرار گرفته و به یک لایه کاملاً‑ متصل داده می‌شوند. خروجی این لایه نیز به یک لایه کاملاً ‑ متصل دیگر داده می‌شود که یک عدد در بازه [۰...۱] تولید می‌کند. عدد تولید شده در این لایه نشان‌دهنده درصدی از ناحیه کاندیدا است که شامل متن است. به عنوان مثال ناحیه‌ای که نصف آن متن بوده و نصف دیگر پس‌زمینه باشد دارای خروجی ۰/۵ خواهد بود. دلیل این کار استفاده از این خروجی در طی روند یادگیری است که در بخش ۴.۶ شرح داده شده است. ساختار شبکه پیشنهادی نیز در شکل ۶.۶ نشان داده شده است.



شکل ۶.۶: معماری شبکه پیشنهادی

پس از اعمال شبکه کانولوشنی روی تصویر ورودی و به دست آمدن خروجی شبکه برای نواحی کاندیدا، از روش حذف غیر بیشینه [۲] [۱۰۸] برای حذف مناطق با هم‌پوشانی بالا استفاده می‌شود. پس از حذف مناطق هم‌پوشان، مناطقی که بالای ۳۰ درصد ناحیه مشترک داشته باشند با هم ادغام می‌شوند که به صورت تجربی به دست آمده است.

[۲] Non Maximum Suppression

۴.۶ نتایج تجربی

برای تولید مجموعه داده‌های آموزشی، الگوریتم تبدیل رنگ لبه‌ها روی مجموعه داده‌های آموزشی اعمال شد. پس از آن مقدار هم‌پوشانی هر یک از نواحی با بخش‌های متنی محاسبه شده و به عنوان خروجی مورد نظر برای آن ناحیه در نظر گرفته شد. همچنین برای نواحی که مقدار هم‌پوشانی آن‌ها با نواحی متنی کمتر از ۵٪ بود، مقدار خروجی ۰ در نظر گرفته شد. این کار برخلاف روند معمول است که مناطق با درصد هم‌پوشانی بالاتر از یک حد آستانه را کلاس مثبت در نظر گرفته و مابقی را کلاس منفی در نظر می‌گیرند؛ استفاده از چنین روشی بخشی از اطلاعات موجود برای آموزش شبکه را حذف می‌کند؛ اما استفاده از میزان هم‌پوشانی به عنوان خروجی به شبکه کمک خواهد کرد تا راحت‌تر بین مناطق متنی و غیرمتنی تفاوت قائل شود، چون به صورت صریح درصد متن موجود در ناحیه کاندیدا به شبکه آموزش داده می‌شود.

در هر یک از تکرارهای الگوریتم آموزش شبکه، برای هر تصویر، با توجه به اینکه تعداد نواحی غیرمتنی معمولاً بیشتر از نواحی متنی است، تمامی نواحی که مقدار خروجی آن‌ها غیر صفر بود انتخاب شده و به تعداد حداکثر دو برابر آن‌ها نواحی با خروجی صفر انتخاب می‌شوند. نحوه انتخاب نواحی با خروجی صفر نیز به این صورت است که نواحی بر حسب خروجی که شبکه در تکرار جاری برای آن‌ها تولید می‌کند مرتب شده و نواحی انتخاب می‌شوند که مقدار خروجی برای آن‌ها بیشتر بوده است. دلیل این کار آن است که خروجی بیشتر نشان‌دهنده خطای بیشتر برای این نواحی است و انتخاب نواحی با خطای بیشتر باعث می‌شود تا در تکرارهای بعدی تمرکز روی این داده‌های با خطای بیشتر افزایش یافته و روند یادگیری تسریع شود.

برای افزایش تعداد نمونه‌های آموزشی، در هر تکرار روی هر یک از تصاویر آموزشی تغییراتی ایجاد شده و تصاویر جدید تولید شده به مجموعه داده‌ها اضافه شدند. تغییرات ایجاد شده شامل برعکس نمودن افقی و عمودی تصاویر ورودی، تغییر کنتراست و چرخش تصویر ورودی هستند. این کار باعث می‌شود تعداد زیادی نمونه آموزشی جدید به مجموعه داده

جدول ۱.۶: نتایج روش پیشنهادی روی مجموعه داده ICDAR 2013

صحت	بازشناسی	F-measure
۹۵/۲۰	۷۹/۵۰	۸۶/۶۴

جدول ۲.۶: نتایج روش پیشنهادی روی مجموعه تصاویر فارست

صحت	بازشناسی	F-measure
۹۳/۵۰	۸۹/۷۳	۹۱/۵۸

اولیه اضافه شوند که باعث بهبود کارایی شبکه می‌شوند.

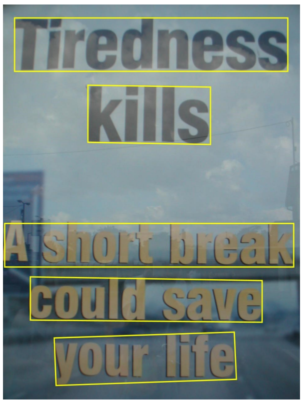
نتایج پیاده‌سازی روش پیشنهادی روی مجموعه داده‌های $ICDAR2013$ و فارست به ترتیب در جدول ۱.۶ و جدول ۲.۶ آورده شده است. همچنین شکل ۷.۶ و شکل ۸.۶ تعدادی از خروجی‌های روش پیشنهادی را به ترتیب برای مجموعه تصاویر $ICDAR2013$ و فارست نشان می‌دهند.

۵.۶ روش مبتنی بر یادگیری عمیق و الگوریتم فرا اکتشافی

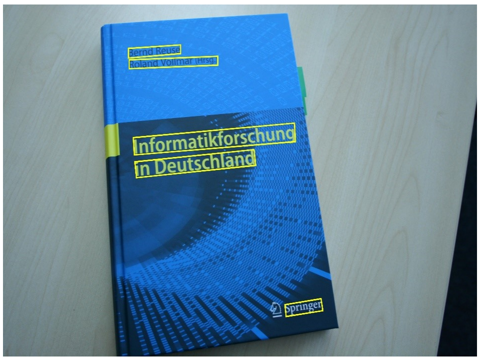
روش دومی که در این فصل ارائه می‌شود از همان ساختار شبکه عصبی پیشنهادی استفاده می‌کند امّا نحوه به دست آوردن نواحی کاندیدا در این روش متفاوت است. ایده اصلی در این روش آن است که الگوریتم با تعدادی ناحیه کاندیدای اولیه شروع کرده و در طی تکرارهای متوالی سعی می‌کند تا این نواحی را بهبود دهد. شکل ۹.۶ یک مثال را در این زمینه نشان می‌دهد. همانگونه که در این شکل مشاهده می‌شود الگوریتم با یک پنجره اولیه شروع کرده و در هر مرحله این پنجره را بهبود داده است تا به پنجره‌ای برسد که بیشتر آن را متن تشکیل می‌دهد؛ با توجه به اینکه خروجی شبکه کانولوشنی معرفی شده در این فصل درصد متن موجود در یک پنجره را نشان می‌دهد می‌تواند به صورت مؤثری به این ایده کمک نماید.

روند کار این روش را می‌توان به دو مرحله تقسیم کرد:

- در ابتدا تصویر ورودی به شبکه کانولوشنی داده می‌شود تا کانال‌های ویژگی برای آن استخراج شود؛ در ادامه، این تصویر را به شبکه‌ای از سلول‌های غیر هم‌پوشان تقسیم کرده و هر یک از این سلول‌ها را به عنوان نواحی کاندیدای اولیه در نظر می‌گیریم.



(آ)



(ب)

(ج)

(د)

(ه)

(و)

شکل ۷.۶: برخی از خروجی‌های روش پیشنهادی برای مجموعه داده $ICDAR2013$

- در مرحله دوم الگوریتم، بخشی از شبکه که پس از لایه‌های ادغام ناحیه مطلوب قرار دارد مورد استفاده قرار می‌گیرد؛ نحوه کار به این صورت است که روی هر یک از سلول‌های ایجاد شده در مرحله قبل، الگوریتم تپه‌نوردی[۳] به صورت جداگانه اعمال

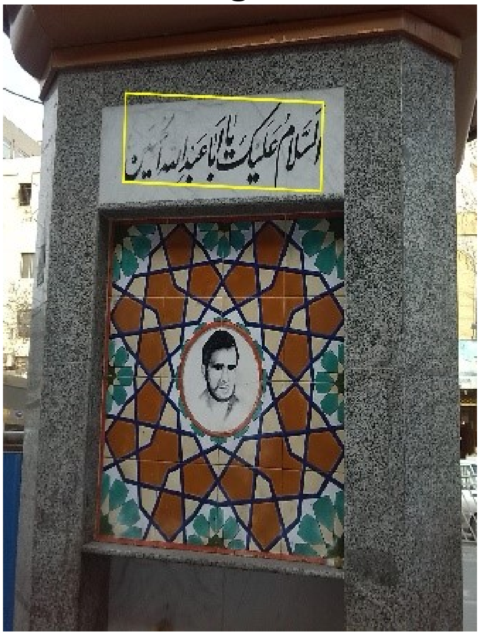
[۳] Hill Climbing

(آ)

(ب)

(ج)

(د)



(ه)

(و)

شکل ۸.۶: برخی از خروجی‌های روش پیشنهادی برای مجموعه داده فارست

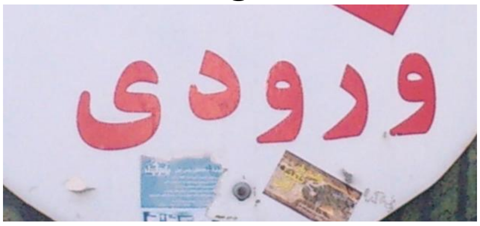
می‌شود؛ برای هر سلول در هر تکرار هشت عمل امکان‌پذیر است که شامل افزایش یا کاهش اندازه پنجره در هر یک از اضلاع سلول است؛ این موضوع در شکل ۱۰.۶

(آ)

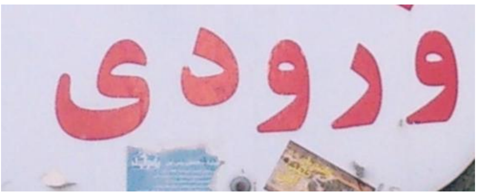
(ب)

(ج)

(د)



(ه)

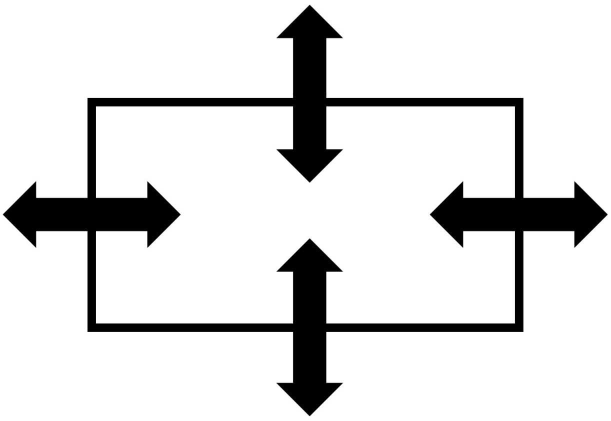


(و)

شکل ۹.۶: روند تدریجی همگرایی به بخش متنی برای یک تصویر نمونه

نشان داده شده است. الگوریتم تپه‌نوردی در هر تکرار، خروجی شبکه را در صورت اعمال هر یک از این عملیات روی سلول به دست می‌آورد و اگر بهترین آن‌ها از خروجی فعلی سلول بهتر باشد آن عمل را روی سلول اعمال می‌کند و این کار را تا زمانی انجام می‌دهد که خروجی سلول بهتر نشود. در نهایت سلول‌هایی که خروجی نهایی آن‌ها کمتر از ۰/۲ باشد حذف می‌شوند.

- دو مرحله قبلی با استفاده از ۳ اندازه سلول متفاوت تکرار می‌شود تا الگوریتم بتواند اندازه‌های مختلف متنی را شناسایی کند. به دلیل آن که معمولاً عرض متون از ارتفاع آن‌ها بیشتر است، عرض این سلول‌ها ۳ برابر ارتفاع آن‌ها در نظر گرفته شده است؛ نحوه کار به این صورت است که ارتفاع تصویر ورودی به ۲، ۶ و ۱۰ بخش تقسیم شده و با توجه به ارتفاع به دست آمده، عرض نیز محاسبه می‌گردد. ناحیه متنی

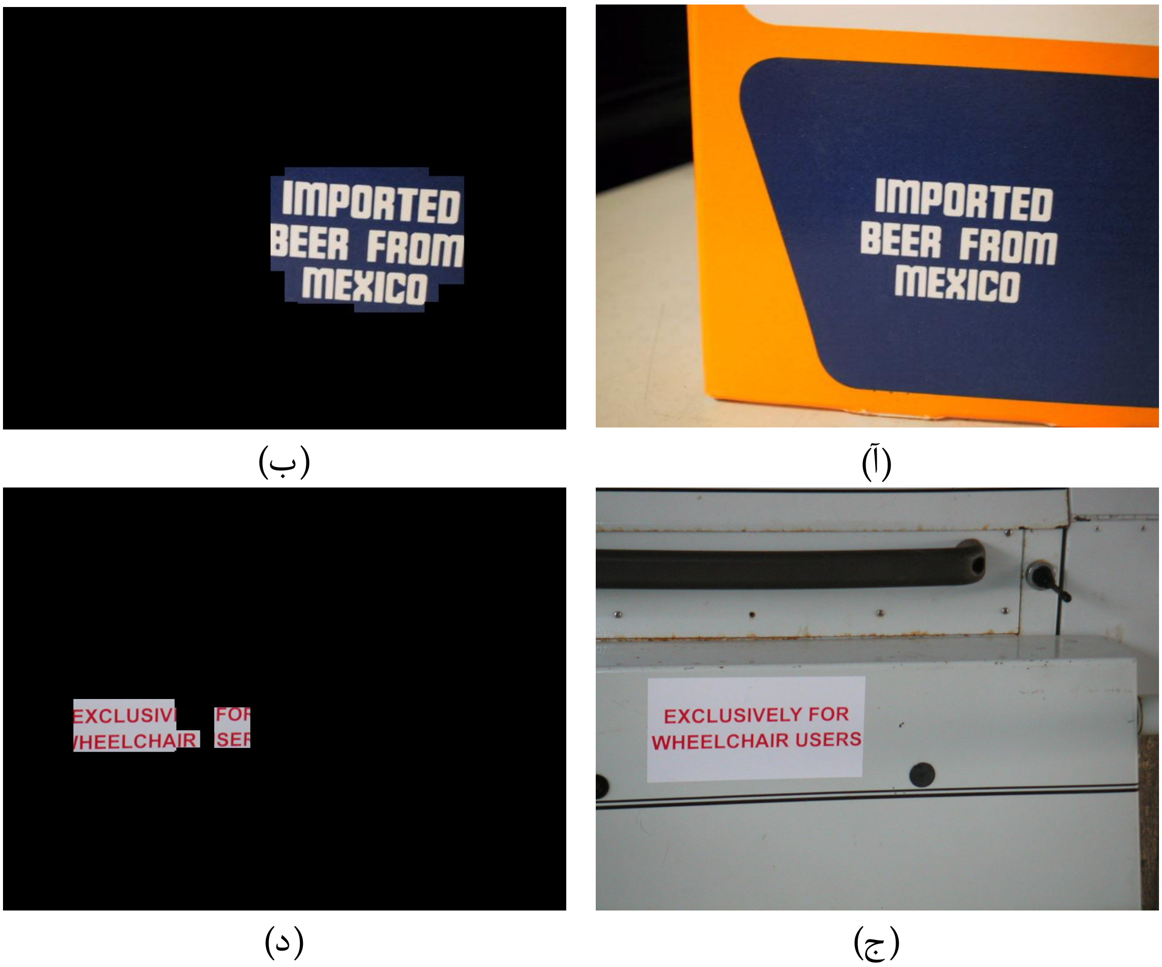


شکل ۱۰.۶: عملیات ممکن برای هر سلول

شناسایی شده توسط دو مرحله قبلی برای دو تصویر نمونه در شکل ۱۱.۶ نشان داده شده است. همانگونه که ملاحضه می‌شود خروجی این الگوریتم کامل نیست؛ در شکل ۱۱.۶ب تمام بخش‌های متنی به صورت یک ناحیه شناسایی شده است و در شکل ۱۱.۶د بخشی‌هایی از متن تشخیص داده نشده است. در ادامه الگوریتم برای بازیابی بخش‌هایی که به اشتباه حذف شده‌اند و نیز جدا کردن خطوط متنی از یکدیگر، لبه‌های بخش‌های متنی شناسایی شده به عنوان نقاط شروع الگوریتم رشد لبه مورد استفاده قرار گرفته و نواحی کاندیدای متنی استخراج شده و دسته‌بندی می‌شوند.

۶.۶ نتایج تجربی

برای تولید داده‌های آموزشی شبکه کانولوشنی در این مرحله، علاوه بر داده‌های که در بخش قبلی برای آموزش شبکه مورد استفاده قرار گرفتند داده‌های دیگری نیز اضافه شدند. نحوه اضافه شدن داده‌ها به این ترتیب است که پنجره‌هایی به صورت تصادفی روی تصاویر ایجاد شده و پس از آن رفتار الگوریتم تپه‌نوردی در صورتی که شبکه به صورت دقیق آموزش داده شده باشد شبیه‌سازی شده و در هر مرحله پنجره‌هایی که الگوریتم تپه‌نوردی ایجاد می‌کند به همراه خروجی مورد نظر برای آن‌ها تولید شده و به مجموعه داده‌های آموزشی شبکه اضافه می‌شوند. بقیه مراحل آموزش شبکه به همان ترتیبی انجام می‌شوند که در بخش ۴.۶ ذکر شد و تنها تفاوت میان این دو روش آن است که خروجی پنجره‌هایی که در این روش



شکل ۱۱.۶: خروجی الگوریتم تپه‌نوردی برای دو تصویر نمونه

جدول ۳.۶: نتایج روش پیشنهادی روی مجموعه داده ICDAR 2013

صحت	بازشناسی	F-measure
۹۴/۹۳	۷۵/۷	۸۴/۲۴

کمتر از ۵٪ هستند صفر نمی‌شود.

نتایج پیاده‌سازی روش پیشنهادی روی مجموعه داده‌های $ICDAR2013$ و فارست به ترتیب در جدول ۱.۶ و جدول ۲.۶ آورده شده است. همچنین شکل ۷.۶ و شکل ۸.۶ تعدادی از خروجی‌های روش پیشنهادی را به ترتیب برای مجموعه تصاویر $ICDAR2013$ و فارست نشان می‌دهند.

جدول ۴.۶: نتایج روش پیشنهادی روی مجموعه تصاویر فارست

صحت	بازشناسی	F-measure
۹۳/۲	۸۴/۶۱	۸۸/۷۰

۷.۶ جمع‌بندی

در این فصل روش‌های «تبدیل رنگ عمیق» و «شبکه فرا اکتشافی» که هر دو مبتنی بر یادگیری عمیق هستند معرفی شدند. تبدیل رنگ عمیق از دو جهت روش تبدیل رنگ لبه‌ها را بهبود می‌دهند؛ بهبود اول در این زمینه صورت گرفته است که با تغییراتی در روش یافتن نواحی کاندیدای متنی، الگوریتم قادر است تا متون با پس‌زمینه ناهموار را نیز تشخیص دهد. همچنین یک کلاسه‌بندی شبکه عصبی کانولوشنی با ساختاری جدید جایگزین کلاسه‌بند آبشاری شده است که کارایی بهتری نسبت به کلاسه‌بند آبشاری دارد. شبکه فرا اکتشافی نیز با به کار گرفتن شبکه کانولوشنی تبدیل رنگ لبه‌ها و استفاده از یک الگوریتم فرا اکتشافی در ابتدا بخش‌های زیادی از مناطق غیر متنی را حذف می‌کند و پس از آن با استفاده از تبدیل رنگ لبه‌ها به شکل دقیق‌تری مناطق متنی را استخراج می‌کند. این روش سریع‌تر از تبدیل رنگ عمیق است امّا دقت کمتری نسبت به آن دارد.

(آ)

(ب)

(ج)

(د)

(ه)

(و)

شکل ۱۲.۶: برخی از خروجی‌های روش پیشنهادی مبتنی بر تپه‌نوردی برای مجموعه داده $ICDAR2013$

شکل ۱۳.۶: برخی از خروجی‌های روش پیشنهادی مبتنی بر الگوریتم تپه‌نوردی برای مجموعه داده فارست

فصل ۷

نتیجه‌گیری و کارهای آینده

۱.۷ مقدمه

در این رساله چهار روش برای مکان‌یابی متون در تصاویر صحنه طبیعی ارائه شد. روش اول که در فصل ۴ ارائه شد با استفاده از امضای رنگ لبه‌ها و الگوریتم خوشه‌بندی اتصال_تکی به تولید نواحی کاندیدای متنی می‌پردازد؛ در نهایت یک کلاسه‌بند آبشاری برای کلاسه‌بند مناطق کاندیدا به متن و غیر متن مورد استفاده قرار می‌گیرد. این روش در فصل ۵ با معرفی الگوریتم «رشد لبه» به همراه «تبدیل رنگ لبه‌ها» به لحاظ پیچیدگی زمانی و دقت عملکرد بهبود داده شد. در فصل ۶ نیز دو روش بر پایه یادگیری عمیق معرفی شدند؛ روش اول با نام تبدیل رنگ عمیق به دو مشکل الگوریتم ارائه شده در فصل ۵ پرداخته و برای هر یک از مشکل‌های مطرح شده به ارائه راه حل می‌پردازد؛ روش شبکه فرا اکتشافی نیز از الگوریتم تپه‌نوردی روی خروجی شبکه کانولوشنی پایه استفاده کرده و سعی می‌کند تا نواحی متنی را شناسایی کند. در این فصل به مقایسه روش‌های ارائه شده با یکدیگر و با چندین روش

جدول ۱.۷: مقایسه روش‌های پیشنهادی روی مجموعه داده ICDAR 2013

روش	صحت	بازشناسی	F-measure
Lu [۴۱]	۸۸/۵۵	۶۹/۵۸	۷۸/۱۹
TexStar [۱۰۵]	۸۸/۴۷	۶۶/۴۵	۷۵/۸۹
TextSpotter [۶۶]	۸۷/۵۱	۶۴/۸۴	۷۴/۴۹
CASIA [۵۱]	۸۴/۷۰	۶۲/۴۵	۷۲/۱۶
Ghanei [۷۸]	۸۸/۵۵	۶۶/۶۰	۷۶/۰۲
FASText [۳۱]	۸۴/۰۰	۶۹/۳۰	۷۶/۸۰
Zheng [۶۷]	۸۹/۵۰	۷۷/۶۳	۸۳/۱۴
Zhao [۱۵]	۹۸/۰۰	۷۲/۰۰	۸۵/۰۰
Zhang [۶۱]	۸۹/۰۷	۸۲/۸۹	۸۵/۸۷
امضای رنگ لبه‌ها	۹۰/۱۰	۷۵/۴۶	۸۲/۱۴
تبدیل رنگ لبه‌ها	۹۳/۸۵	۷۷/۷۷	۸۵/۰۲
تبدیل رنگ عمیق	۹۵/۲۰	۷۹/۵۰	۸۶/۶۴
شبکه فرا اکتشافی	۹۴/۹۳	۷۵/۷۰	۸۴/۲۴

دیگر می‌پردازیم.

۲.۷ مقایسه روش‌های پیشنهادی

برای پیاده‌سازی روش‌های ارائه شده، محیط زبان‌های متلب و پایتون به همراه جعبه‌ابزار یادگیری عمیق $MXNet$ [۱۰۹] مورد استفاده قرار گرفتند. سیستم مورد استفاده برای پیاده‌سازی روش‌های پیشنهادی دارای پردازه اینتل $i7-7700K$ به همراه ۱۶ گیگابایت حافظه و واحد پردازش گرافیکی $Nvidia\ GeForce\ GTX\ 1080$ است. نتایج پیاده‌سازی روش‌های پیشنهادی به همراه خروجی چندین روش دیگر برای مجموعه داده‌های انگلیسی و فارسی به ترتیب در جدول‌های ۱.۷ و ۲.۷ آورده شده است. برای تصاویر فارسی از کد به اشتراک گذاشته شده توسط خود نویسندگان مقاله استفاده شده است. همانگونه که مشاهده می‌شود روش تبدیل رنگ عمیق که از ترکیب یادگیری عمیق و «الگوریتم رشد لبه» برای مکان‌یابی متون استفاده می‌کند از لحاظ معیار F-measure بهتر از روش‌های دیگر عمل کرده است و روش‌های تبدیل رنگ لبه‌ها، شبکه فرا اکتشافی و امضای رنگ لبه‌ها در رتبه‌های بعدی قرار دارند.

جدول ۲.۷: مقایسه روش‌های پیشنهادی روی مجموعه داده فارست

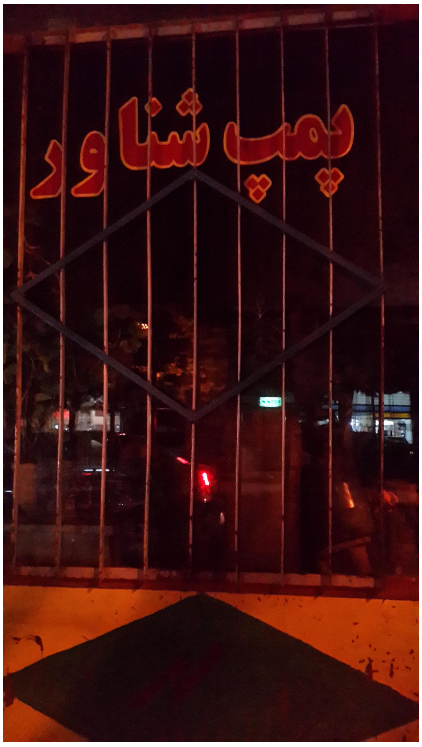
روش	صحت	بازشناسی	F-measure
Yao [۵۴]	۶۳/۵۶	۴۷/۸۵	۵۴/۶۰
Neumann [۶۵]	۴۸/۱۱	۴۱/۶۲	۴۴/۶۳
امضای رنگ لبه‌ها	۹۰/۸۴	۸۷/۵۴	۸۹/۱۶
تبدیل رنگ لبه‌ها	۹۲/۰۲	۸۸/۱۶	۹۰/۰۵
تبدیل رنگ عمیق	۹۳/۵۰	۸۹/۷۳	۹۱/۵۸
شبکه فرا اکتشافی	۹۳/۲۰	۸۴/۶۱	۸۸/۷۰

۳.۷ مواردی که متن به درستی تشخیص داده نمی‌شود

روش امضای رنگ لبه‌ها گاهی اوقات در تصاویری که لبه‌های متنی آن‌ها با لبه‌های غیر متنی تداخل پیدا می‌کنند دچار مشکل می‌شود. نمونه‌هایی از تصاویری که روش ارائه شده در فصل ۴ قادر به تشخیص آن‌ها نبوده است در شکل ۱.۷ نشان داده شده است. همانگونه که مشاهده می‌شود به دلیل تداخلی که میان پیکسل‌های لبه نرده‌ها با پیکسل‌های لبه بخش متنی وجود داشته است و با توجه به اینکه هم بخش متنی و هم بخش غیر متنی دارای پس‌زمینه یکسانی هستند، این بخش‌ها در هم ادغام شده و مانع ایجاد ناحیه کاندیدای مناسب می‌شوند.

روش تبدیل رنگ لبه‌ها نیز در مواردی که بخشی از پس‌زمینه روشن‌تر از متن بوده و بخشی دیگر تیره‌تر از آن باشد، یا در مواردی که پس‌زمینه پیچیدگی زیادی داشته باشد قادر به تشخیص بخش متنی نیست. چند نمونه از مواردی که این روش قادر به تشخیص نواحی متنی نبوده است در شکل ۲.۷ نشان داده شده است. در تصویر ۲.۷آ کلاسه‌بند قادر نبوده است تا کلمه «همشهری» را به عنوان متن تشخیص دهد. همچنین در تصویر ۲.۷ب کلمه $wolfgang$ به دلیل قرابت رنگی لبه‌های اطراف و عدم تشکیل ناحیه کاندیدای متنی برای این کلمه تشخیص داده نشده است.

مشکل مربوط به کلاسه‌بند در تصویر ۲.۷آ با استفاده از شبکه عصبی عمیق در روش تبدیل رنگ عمیق که در فصل ۶ ارائه شد رفع شده است اما مشکل مربوط تصویر دوم همچنان در روشی که از یادگیری عمیق نیز استفاده می‌کند وجود دارد؛ چرا که هر دو روش از مولد

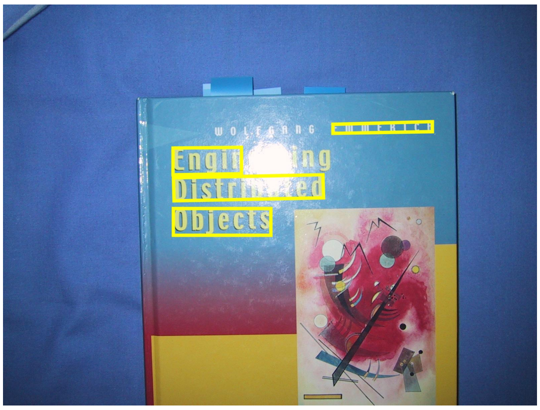


(آ)

(ب)

شکل ۱.۷: برخی از مواردی که روش امضای رنگ لبه‌ها قادر به تشخیص متن نیست

(آ)



(ب)

شکل ۲.۷: برخی از مواردی که روش تبدیل رنگ لبه‌ها قادر به تشخیص متن نیست

یکسانی برای تولید نواحی کاندیدای متنی استفاده می‌کنند.

در نهایت آخرین روش ارائه شده به دلیل آن که از یک سری پنجره اولیه شروع کرده و در نهایت به بخش‌های متنی همگرا می‌شود در برخی موارد قادر نیست تا تمامی بخش‌های متنی را تشخیص دهد. نمونه‌هایی از این موارد در شکل ۳.۷ نشان داده شده است. همچنین هیچ یک از روش‌های پیشنهادی قادر نیستند متونی را تشخیص دهند که تنها شامل یک

(آ)

(ب)

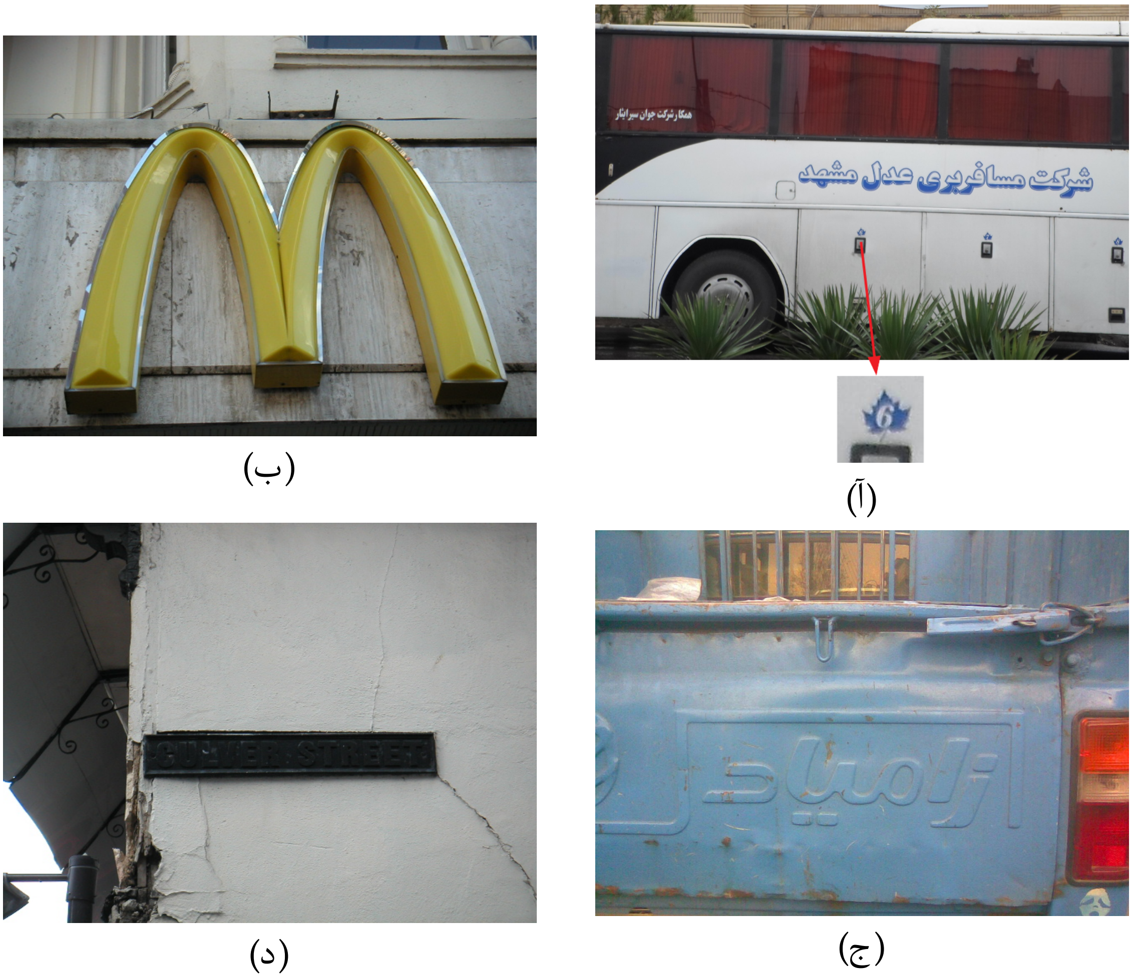
شکل ۳.۷: برخی از مواردی که روش شبکه فرا اکتشافی قادر به تشخیص متن نیست. بخش‌های متنی تشخیص داده نشده برای مشاهده بهتر در بالای هر تصویر به صورت بریده شده نشان داده شده‌اند

حرف یا یک رقم هستند؛ همچنین در شرایطی که کنتراست میان متن و پس‌زمینه بسیار پایین باشد نیز هیچ یک از روش‌ها قادر به تشخیص متن نیستند. نمونه‌هایی از این تصاویر نیز در شکل ۴.۷ نشان داده شده است.

۴.۷ مواردی از تشخیص اشتباه مناطق غیر متنی

در برخی موارد نشانه‌گذاری ناقص مجموعه تصاویر باعث ایجاد برخی از اشتباهات شده است. به عنوان مثال یک نمونه تصویر در شکل ۵.۷آ نشان داده شده است که ناحیه تشخیص داده شده در آن متن است امّا به دلیل نشانه‌گذاری نشدن، این ناحیه به عنوان تشخیص اشتباه در نظر گرفته می‌شود.

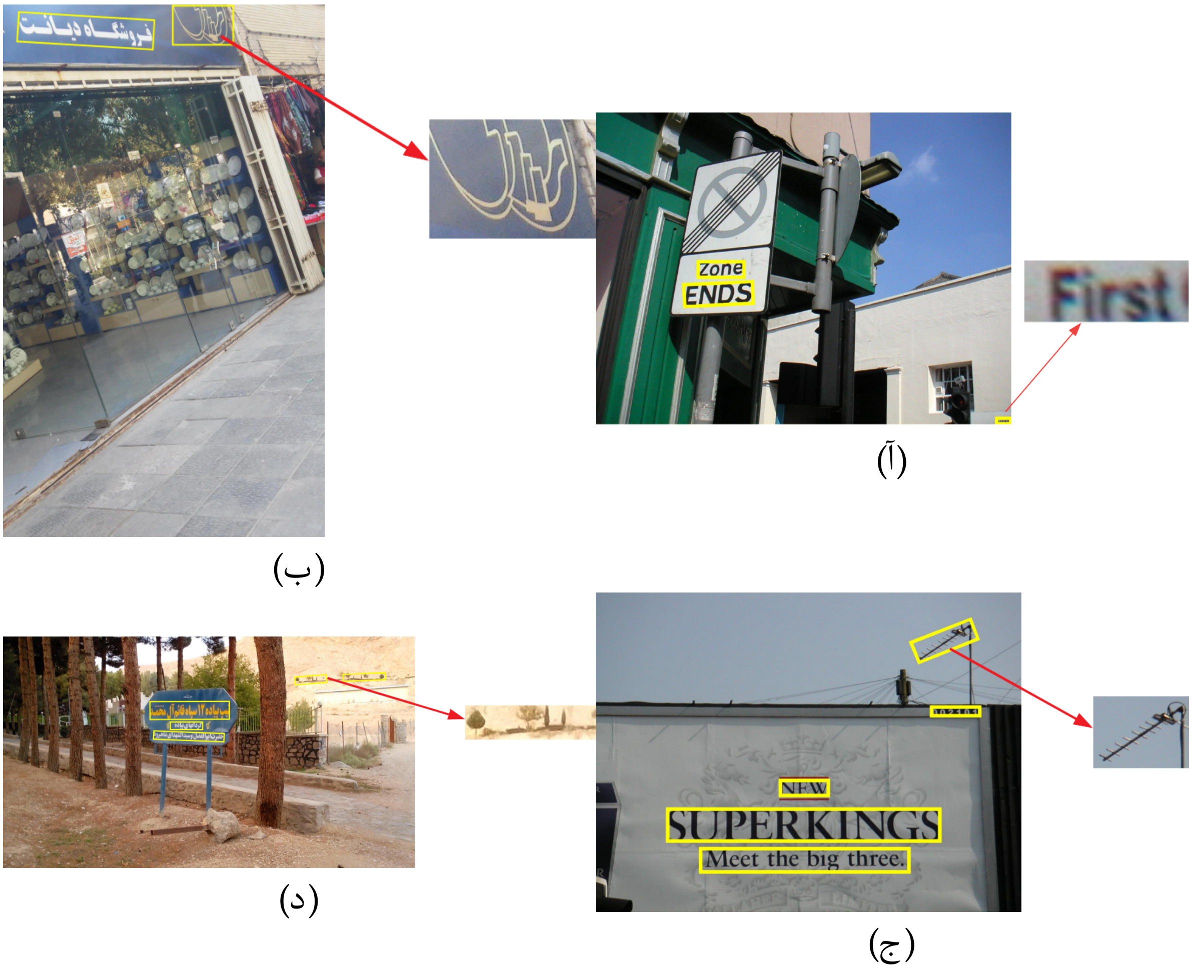
یکی دیگر از مواردی که به اشتباه متن تشخیص داده می‌شوند، اشیایی هستند که توسط



(آ) (ب)
(ج) (د)

شکل ۴.۷: برخی از تصاویری که هیچ یک از روش‌های پیشنهادی قادر به تشخیص متن نبوده‌اند. ردیف اول برای متون با یک کاراکتر و ردیف دوم برای تصاویر با کنتراست بسیار کم

انسان ایجاد شده و شکل منظم و شبیه به متن دارند. دو مورد از این اشتباهات نیز در شکل ۵.۷ب و ۵.۷ج به نمایش در آمده است. همانگونه که مشاهده می‌شود ناحیه‌ای که به اشتباه تشخیص داده شده است شباهت زیادی به متن فارسی دارد. گیاهان نیز گاهی اوقات به اشتباه به عنوان متن تشخیص داده می‌شوند که یک مورد در شکل۵.۷د نشان داده شده است. همچنین نقاط ضعف و قوت هر یک از روش‌های ارائه شده در جدول ۳.۷ خلاصه شده است. نکته مهمی که در مورد تمامی روش‌های ارائه شده وجود دارد عدم وجود پیش‌فرض در مورد ساختار فونت‌ها و اندازه آن‌هاست. این مورد باعث می‌شود که این روش‌ها بتوانند ساختار پیچیده زبان فارسی را به خوبی مدیریت کنند؛ همچنین به دلیل استفاده از الگوریتم‌های گروه‌بندی لبه‌ها، با توجه به این که نقطه‌های کلمات با بدنه اصلی متن هم‌رنگ هستند روش‌های ارائه شده در بیشتر اوقات قادر هستند تا نقطه‌های



شکل ۵.۷: برخی از نواحی که به دلیل شباهت به متن به اشتباه متن تشخیص داده شده‌اند

نویسه‌های فارسی را نیز با آن‌ها گروه‌بندی کنند؛ این مورد در مورد زبان فارسی مزیت بزرگی به شمار می‌رود چرا که نقطه‌ها اهمیت بسیار زیادی برای درک متن فارسی دارند؛ این در حالی است که روش‌های مبتنی بر تحلیل مؤلفه متصل، نقاط را به دلیل اندازه کاملاً متفاوت آن‌ها نسبت به بدنه اصلی متن به عنوان نویز حذف می‌کنند.

با توجه به این که روش‌های امضای رنگ لبه‌ها و تبدیل رنگ لبه‌ها روی پردازنده مرکزی اجرا می‌شوند امّا دو روش دیگر از کارت گرافیکی برای یافتن مناطق متنی استفاده می‌کنند مقایسه این روش‌ها به لحاظ زمانی منصفانه نیست؛ امّا برای شفافیت بیشتر موضوع، زمان اجرای هر یک از الگوریتم‌ها در این بخش گزارش می‌شود. این اعداد به صورت متوسط گزارش می‌شوند و وضوح متوسط تصاویری که برای به دست آوردن زمان اجرا به کار گرفته شده‌اند ۱۹۳۰ × ۱۶۰۰ بوده است. روش امضای رنگ لبه‌ها برای هر تصویر ورودی تقریباً

به طور متوسط نیاز به ۴۱ ثانیه زمان دارد در حالی که در الگوریتم تبدیل رنگ لبه‌ها این زمان به ۸ ثانیه کاهش پیدا می‌کند. تبدیل رنگ عمیق و شبکه فرا اکتشافی در صورتی که از کارت گرافیکی برای پردازش استفاده کنند برای هر تصویر به ترتیب نیاز به ۶ و ۲ ثانیه زمان دارند؛ امّا در صورتی که از کارت گرافیکی استفاده نکنند با توجه به محاسبات بسیار زیاد مورد نیاز در شبکه‌های کانولوشنی این اعداد به ترتیب با افزایش ۲۵ و ۳۵ برابری رو به رو می‌شوند. نکته مهم در مورد روش شبکه فرا اکتشافی، زمان اجرای کمتر آن نسبت به تبدیل رنگ عمیق علارغم استفاده آن از الگوریتم فرا اکتشافی است. کاهش زمان اجرای این روش به این دلیل اتفاق افتاده است که پردازش‌های مربوط به الگوریتم فرا اکتشافی با استفاده از کارت گرافیکی انجام می‌شوند و باعث می‌شوند تعداد مناطقی که نیاز به استفاده از الگوریتم رشد لبه در آن‌ها وجود دارد به شدت کاهش پیدا کرده و بنابراین زمان پردازش کاهش پیدا کند.

۵.۷ پیشنهاد برای کارهای آینده

تمامی الگوریتم‌های معرفی شده در این رساله وابسته به استخراج صحیح پیکسل‌های لبه تصویر ورودی هستند و در صورتی که در یک تصویر، لبه‌ها به درستی استخراج نشوند مراحل بعدی این روش‌ها دچار مشکل خواهد شد. همچنین برای به دست آوردن رنگ‌های اطراف لبه‌ها، الگوریتم تبدیل رنگ لبه‌ها معرفی شد؛ یکی از نقاط ضعف این الگوریتم در مواردی رخ می‌دهد که پس‌زمینه بسیار ناهموار باشد. این ضعف از آنجا ناشی می‌شود که الگوریتم برای یافتن رنگ‌های اطراف لبه‌ها وابسته به جهت گرادیان در محل لبه‌هاست. برای رفع این مشکلات می‌توان مراحل یافتن لبه‌ها و رنگ‌های اطراف لبه‌ها را با استفاده از شبکه‌های کانولوشنی انجام داد. با توجه به این که هم اکنون نیز بهترین روش ارائه شده در این رساله از یک شبکه کانولوشنی استفاده می‌کند می‌توان از ویژگی‌های استخراج شده از این شبکه برای این منظور استفاده کرد تا هم بار محاسباتی این کار کاهش پیدا کند و هم کارایی این روش‌ها افزایش پیدا کند.

جدول ۳.۷: مقایسه روش‌های پیشنهادی با همدیگر و نقاط و ضعف و قوت آن‌ها

روش	نقاط قوت	نقاط ضعف
امضای رنگ لبه‌ها	۱. عدم محدودیت به فونت و ساختار خاص یک زبان ۲. توانایی شناسایی متون فارسی و انگلیسی ۳. عدم نیاز به کارت گرافیکی برای انجام محاسبات در زمان آموزش و آزمون	۱. عدم تشکیل نواحی کاندیدا در مناطقی که لبه‌های متنی و غیر متنی تداخل دارند ۲. مرتبه زمانی $O(n^2)$ الگوریتم خوشه بندی به کار رفته در آن
تبدیل رنگ لبه‌ها	۱. توانایی تشکیل نواحی کاندیدای متنی در مناطقی که بین لبه‌های متنی و غیر متنی تداخل وجود دارد ۲. الگوریتم تشکل لبه‌های متنی که دارای پیچیدگی زمانی $O(n)$ است ۳. عدم نیاز به کارت گرافیکی برای انجام محاسبات در زمان آموزش و آزمون	۱. عدم توانایی ایجاد نواحی کاندیدا در مناطق با پس‌زمینه ناهموار
تبدیل رنگ عمیق	۱. توانایی تشکیل نواحی کاندیدا در مناطق با پس‌زمینه ناهموار ۲. استفاده از شبکه عصبی کانولوشنی برای کلاسه‌بندی نواحی کاندیدا با دقت بیشتر نسبت به کلاسه‌بند آبشاری	۱. نیاز به کارت گرافیکی برای آموزش و آزمون
شبکه فرا اکتشافی	۱. استفاده از الگوریتم فرا اکتشافی برای کاهش مناطقی که باید از آن‌ها نواحی کاندیدا استخراج شود ۲. سرعت بیشتر نسبت به روش تبدیل رنگ عمیق	۱. نیاز به کارت گرافیکی برای آموزش و آزمون ۲. دقت کمتر نسبت به تبدیل رنگ عمیق

یکی دیگر از ضعف‌های روش پیشنهادی عدم قابلیت آن برای مکان‌یابی متونی است که به صورت عمودی یا نزدیک به عمودی قرار دارند. برای رفع این مشکل می‌توان در الگوریتم رشد لبه به جای یک محدوده همسایگی از چندین محدوده استفاده کرد که هر یک از آن‌ها وظیفه یافتن متن در جهت‌های مختلف را به عهده خواهند داشت. امّا در بخش کلاسه‌بندی می‌توان از لایه‌های یادگیری عمیق جدیدتری مانند [۱۱۰، ۱۱۱] استفاده کرد که زاویه مناطق کاندیدا را نیز در نظر می‌گیرند.
یکی دیگر از مواردی که می‌تواند در کارهای آتی مورد توجه قرار گیرد تولید مجموعه تصاویر تصنعی صحنه طبیعی فارسی برای بهبود عملکرد روش‌های مبتنی بر یادگیری عمیق است؛ در صورتی که مجموعه تصاویر تصنعی فارسی برای چنین مواردی تولید شود نه تنها به بهبود الگوریتم‌های موجود کمک شایانی خواهد کرد بلکه نیاز به نشانه‌گذاری دستی برای تصاویر را نیز مرتفع می‌کند. ضمن اینکه با توجه به خاصیت ترتیبی متن می‌توان از لایه‌هایی مانند $LSTM$در شبکه استفاده کرد که قابلیت در نظر گرفتن ترتیب کاراکترها در متن را دارند. همچنین در چنین مواردی با توجه به اینکه متن نوشته شده در تصویر در دسترس است می‌توان سیستم‌هایی را آموزش داد که همزمان هم می‌توانند متن موجود در تصاویر را تشخیص دهند و هم آن‌ها را بازخوانی کنند.
نکته مهم دیگری که در این زمینه وجود دارد پژوهش در زمینه بازشناسی متون فارسی در تصاویر صحنه طبیعی است؛ در این زمینه در مورد زبان انگلیسی پژوهش‌های زیادی انجام گرفته است امّا در مورد زبان فارسی کار زیادی انجام نشده است و زمینه مناسبی برای پژوهش در این مورد وجود دارد. این سیستم‌ها در صورتی که از مدل‌های زبانی در ایجاد آن‌ها کمک گرفته شود می‌توانند در حذف مناطقی که به اشتباه متن تشخیص داده شده‌اند نیز بسیار مؤثر باشند.

مراجع

[1] F. Chollet. *Deep Learning with Python*. Shelter Island, New York: Manning Publications, 1st edition ed. , Dec. 2017.

[2] R. Girshick, “Fast R-CNN,” in *2015 IEEE International Conference on Computer Vision (ICCV)*, (Santiago, Chile), pp.1440–1448, Dec. 2015.

[3] H. Zhang, K. Zhao, Y.-Z. Song, and J. Guo, “Text extraction from natural scene image: A survey,” *Neurocomputing*, vol.122, pp.310–323, Dec. 2013.

[4] L. Neumann and J. Matas, “A Method for Text Localization and Recognition in Real-World Images,” in *Computer Vision – ACCV 2010*, (Berlin, Heidelberg), pp.770–783, Springer, Nov. 2010.

[5] X. Chen and A. L. Yuille, “Detecting and reading text in natural scenes,” in *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol.2, pp.359–366, IEEE, 2004.

[6] Y. Pan, X. Hou, and C. Liu, “Text Localization in Natural Scene Images Based on Conditional Random Field,” in *2009 10th International Conference on Document Analysis and Recognition*, pp.6–10, July 2009.

[7] J. Sochman and J. Matas, “Waldboost-learning for time constrained sequential detec-

tion," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol.2, pp.150–156, IEEE, 2005.

[8] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol.1, pp.886–893, IEEE, 2005.

[9] W. Niblack. *An introduction to digital image processing*. Birkeroed, Denmark: Strandberg Publishing Company, 1985.

[10] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proceedings of the eighteenth international conference on machine learning, ICML*, vol.1, pp.282–289, 2001.

[11] J.-J. Lee, P.-H. Lee, S.-W. Lee, A. Yuille, and C. Koch, "AdaBoost for Text Detection in Natural Scene," in *2011 International Conference on Document Analysis and Recognition*, pp.429–434, IEEE, Sept. 2011.

[12] R. E. Schapire and Y. Singer, "Improved boosting algorithms using confidence-rated predictions," *Machine learning*, vol.37, no.3, pp.297–336, 1999.

[13] Y.-F. Pan, X. Hou, and C.-L. Liu, "A Robust System to Detect and Localize Texts in Natural Scene Images," in *The Eighth IAPR International Workshop on Document Analysis Systems*, pp.35–42, IEEE, Sept. 2008.

[14] R. Minetto, N. Thome, M. Cord, N. J. Leite, and J. Stolfi, "T-HOG: An effective gradient-based descriptor for single line text regions," *Pattern Recognition*, vol.46, pp.1078–1090, Mar. 2013.

[15] F. Zhao, Y. Yang, H. Zhang, L. Yang, and L. Zhang, "Sign text detection in street

view images using an integrated feature," *Multimedia Tools and Applications*, vol.77, pp.28049–28076, Nov. 2018.

[16] S. Leutenegger, M. Chli, and R. Y. Siegwart, "BRISK: Binary Robust invariant scalable keypoints," in *2011 International Conference on Computer Vision*, pp.2548–2555, Nov. 2011.

[17] A. Atsalakis and N. Papamarkos, "Color reduction and estimation of the number of dominant colors by using a self-growing and self-organized neural gas," *Engineering Applications of Artificial Intelligence*, vol.19, pp.769–786, Oct. 2006.

[18] S. Tian, Y. Pan, C. Huang, S. Lu, K. Yu, and C. Lim Tan, "Text flow: A unified text detection system in natural scene images," in *The IEEE international conference on computer vision (ICCV)*, pp.4651–4659, Dec. 2015.

[19] Z. Tian, W. Huang, T. He, P. He, and Y. Qiao, "Detecting Text in Natural Image with Connectionist Text Proposal Network," in *Computer Vision – ECCV 2016*, pp.56–72, Springer, 2016.

[20] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional LSTM and other neural network architectures," *Neural Networks*, vol.18, pp.602–610, July 2005.

[21] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International conference on learning representations*, 2015.

[22] Z. Zhong, L. Jin, S. Zhang, and Z. Feng, "Deeptext: A unified framework for text proposal generation and text detection in natural images," *CoRR*, vol.abs/1605.07314, 2016.

[23] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke,

and A. Rabinovich, “Going deeper with convolutions,” in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[24] J. Ohya, A. Shio, and S. Akamatsu, “Recognizing characters in scene images,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.16, pp.214–220, Feb. 1994.

[25] C. Li, X. Ding, and Y. Wu, “Automatic text location in natural scene images,” in *Proceedings of Sixth International Conference on Document Analysis and Recognition*, pp.1069–1073, Sept. 2001.

[26] K. C. Kim, H. R. Byun, Y. J. Song, Y.-W. Choi, S. Y. Chi, K. K. Kim, and Y. Chung, “Scene text extraction in natural scene images using hierarchical feature combining and verification,” in *Proceedings of the 17th International Conference on Pattern Recognition*, vol.2, pp.679–682, IEEE, 2004.

[27] H. Takahashi and M. Nakajima, “Region graph based text extraction from outdoor images,” in *Third International Conference on Information Technology and Applications (ICITA'05)*, vol.1, pp.680–685, IEEE, 2005.

[28] B. Epshtein, E. Ofek, and Y. Wexler, “Detecting text in natural scenes with stroke width transform,” in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp.2963–2970, June 2010.

[29] Y. Li and H. Lu, “Scene text detection via stroke width,” in *21st International Conference on Pattern Recognition*, pp.681–684, IEEE, 2012.

[30] A. Risnumawan, P. Shivakumara, C. S. Chan, and C. L. Tan, “A robust arbitrary text detection system for natural scene images,” *Expert Systems with Applications*, vol.41, pp.8027–8048, Dec. 2014.

[31] M. Buta, L. Neumann, and J. Matas, “FASText: Efficient Unconstrained Scene Text Detector,” in *2015 IEEE International Conference on Computer Vision (ICCV)*, pp.1206–1214, Dec. 2015.

[32] E. Rosten, R. Porter, and T. Drummond, “Faster and Better: A Machine Learning Approach to Corner Detection,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.32, pp.105–119, Jan. 2010.

[33] J. Friedman, T. Hastie, and R. Tibshirani, “Additive logistic regression: a statistical view of boosting (With discussion and a rejoinder by the authors),” *The Annals of Statistics*, vol.28, pp.337–407, Apr. 2000.

[34] V. Goel, A. Mishra, K. Alahari, and C. V. Jawahar, “Whole is greater than sum of parts: Recognizing scene text words,” in *12th International Conference on Document Analysis and Recognition*, pp.398–402, IEEE, 2013.

[35] A. Coates, B. Carpenter, C. Case, S. Satheesh, B. Suresh, T. Wang, D. J. Wu, and A. Y. Ng, “Text Detection and Character Recognition in Scene Images with Unsupervised Feature Learning,” in *2011 International Conference on Document Analysis and Recognition (ICDAR)*, pp.440–445, IEEE, Sept. 2011.

[36] J. Yan and X. Gao, “Detection and recognition of text superimposed in images base on layered method,” *Neurocomputing*, vol.134, pp.3–14, June 2014.

[37] Yi-Feng Pan, Xinwen Hou, and Cheng-Lin Liu, “A Hybrid Approach to Detect and Localize Texts in Natural Scene Images,” *IEEE Transactions on Image Processing*, vol.20, pp.800–813, Mar. 2011.

[38] X. Liu and W. Wang, “An effective graph-cut scene text localization with embedded text segmentation,” *Multimedia Tools and Applications*, vol.74, pp.4891–4906, July 2015.

[39] J.-L. Yao, Y.-Q. Wang, L.-B. Weng, and Y.-P. Yang, “Locating text based on connected component and SVM,” in *2007 International Conference on Wavelet Analysis and Pattern Recognition*, vol.3, pp.1418–1423, IEEE, 2007.

[40] J. Fabrizio, B. Marcotegui, and M. Cord, “Text detection in street level images,” *Pattern Analysis and Applications*, vol.16, no.4, pp.519–533, 2013.

[41] S. Lu, T. Chen, S. Tian, J.-H. Lim, and C.-L. Tan, “Scene text extraction based on edges and support vector regression,” *International Journal on Document Analysis and Recognition (IJDAR)*, vol.18, pp.125–135, June 2015.

[42] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman, “Reading text in the wild with convolutional neural networks,” *International Journal of Computer Vision*, vol.116, no.1, pp.1–20, 2016.

[43] C. L. Zitnick and P. Dollár, “Edge Boxes: Locating Object Proposals from Edges,” in *Computer Vision – ECCV 2014*, Lecture Notes in Computer Science, pp.391–405, Springer, Sept. 2014.

[44] P. Dollár, R. Appel, S. Belongie, and P. Perona, “Fast Feature Pyramids for Object Detection,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.36, pp.1532–1545, Aug. 2014.

[45] L. Breiman, “Random Forests,” *Machine Learning*, vol.45, pp.5–32, Oct. 2001.

[46] X. He, Y. Song, and Y. Zhang, “A coarse-to-fine scene text detection method based on Skeleton-cut detector and Binary-Tree-Search based rectification,” *Pattern Recognition Letters*, vol.112, pp.27–33, Sept. 2018.

[47] X. Huang, “Automatic video scene text detection based on saliency edge map,” *Multimedia Tools and Applications*, vol.78, pp.34819–34838, Dec. 2019.

[48] M. Figueiredo and A. Jain, “Unsupervised learning of finite mixture models,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.24, pp.381–396, Mar. 2002.

[49] D. Deng, H. Liu, X. Li, and D. Cai, “PixelLink: Detecting Scene Text via Instance Segmentation,” in *Thirty-Second AAAI Conference on Artificial Intelligence*, pp.6776–6780, Apr. 2018.

[50] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, “SSD: Single Shot MultiBox Detector,” in *Computer Vision – ECCV 2016*, pp.21–37, Springer International Publishing, 2016.

[51] C. Shi, C. Wang, B. Xiao, Y. Zhang, and S. Gao, “Scene text detection using graph model built upon maximally stable extremal regions,” *Pattern Recognition Letters*, vol.34, pp.107–116, Jan. 2013.

[52] J. Matas, O. Chum, M. Urban, and T. Pajdla, “Robust wide-baseline stereo from maximally stable extremal regions,” *Image and vision computing*, vol.22, no.10, pp.761–767, 2004.

[53] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. V. Gool, “A Comparison of Affine Region Detectors,” *International Journal of Computer Vision*, vol.65, pp.43–72, Nov. 2005.

[54] Yao Li, Wenjing Jia, Chunhua Shen, and A. van den Hengel, “Characterness: An Indicator of Text in the Wild,” *IEEE Transactions on Image Processing*, vol.23, pp.1666–1677, Apr. 2014.

[55] X.-C. Yin, X. Yin, K. Huang, and H.-W. Hao, “Robust text detection in natural scene

images," *IEEE transactions on pattern analysis and machine intelligence*, vol.36, no.5, pp.970–983, 2014.

[56] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data Clustering: A Review," *ACM Computing Surveys*, vol.31, pp.264–323, Sept. 1999.

[57] X.-C. Yin, W.-Y. Pei, J. Zhang, and H.-W. Hao, "Multi-Orientation Scene Text Detection with Adaptive Clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.37, pp.1930–1937, Sept. 2015.

[58] J. Dai, Z. Wang, X. Zhao, and S. Shao, "Scene text detection based on enhanced multi-channels MSER and a fast text grouping process," in *2018 IEEE 3rd International Conference on Cloud Computing and Big Data Analysis (ICCCBDA)*, pp.351–355, Apr. 2018.

[59] R. Baran, P. Partila, and R. Wilk, "Automated Text Detection and Character Recognition in Natural Scenes Based on Local Image Features and Contour Processing Techniques," in *Intelligent Human Systems Integration*, pp.42–48, Springer International Publishing, 2018.

[60] Y. Wang, C. Shi, B. Xiao, C. Wang, and C. Qi, "CRF based text detection for natural scene images using convolutional neural network and context information," *Neurocomputing*, vol.295, pp.46–58, June 2018.

[61] X. Zhang, X. Gao, and C. Tian, "Text detection in natural scene images based on color prior guided MSER," *Neurocomputing*, vol.307, pp.61–71, Sept. 2018.

[62] R. Soni, B. Kumar, and S. Chand, "Text detection and localization in natural scene images based on text awareness score," *Applied Intelligence*, vol.49, pp.1376–1405, Apr. 2019.

[63] K. He and J. Sun, "Fast Guided Filter," *arXiv:1505.00996 [cs]*, May 2015.

[64] E. Alpaydin. *Introduction to Machine Learning*. MIT Press, Dec. 2009. Google-Books-ID: TtrxCwAAQBAJ.

[65] L. Neumann and J. Matas, "Real-time scene text localization and recognition," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp.3538–3545, June 2012.

[66] L. Neumann and J. Matas, "On combining multiple segmentations in scene text recognition," in *2013 12th International Conference on Document Analysis and Recognition*, pp.523–527, IEEE, 2013.

[67] Y. Zheng, Q. Li, J. Liu, H. Liu, G. Li, and S. Zhang, "A cascaded method for text detection in natural scene images," *Neurocomputing*, vol.238, pp.307–315, May 2017.

[68] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[69] A. Gupta, A. Vedaldi, and A. Zisserman, "Synthetic data for text localisation in natural images," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[70] Z. Zhang, C. Zhang, W. Shen, C. Yao, W. Liu, and X. Bai, "Multi-oriented text detection with fully convolutional networks," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[71] S. Zhang, M. Lin, T. Chen, L. Jin, and L. Lin, "Character proposal network for robust text extraction," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp.2633–2637, 2016.

[72] W. He, X.-Y. Zhang, F. Yin, and C.-L. Liu, "Deep direct regression for multi-oriented scene text detection," in *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[73] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[74] D. Bazazian, R. Gómez, A. Nicolaou, L. Gómez, D. Karatzas, and A. D. Bagdanov, "Fast: Facilitated and accurate scene text proposals through fcn guided pruning," *Pattern Recognition Letters*, vol.119, pp.112 – 120, 2019.

[75] L. Gómez and D. Karatzas, "Textproposals: A text-specific selective search algorithm for word spotting in the wild," *Pattern Recognition*, vol.70, pp.60 – 74, 2017.

[76] M. Darab and M. Rahmati, "A Hybrid Approach to Localize Farsi Text in Natural Scene Images," *Procedia Computer Science*, vol.13, pp.171–184, 2012.

[77] S. Ghanei and K. Faez, "Robust Localization of Texts in Real-World Images," *International Journal of Pattern Recognition and Artificial Intelligence*, vol.29, pp.1–26, Nov. 2015.

[78] S. Ghanei and K. Faez, "Localizing scene texts by fuzzy inference systems and low rank matrix recovery model," *Computer Vision and Image Understanding*, vol.142, pp.94–110, Jan. 2016.

[79] X. Shen and Y. Wu, "A unified approach to salient object detection via low rank matrix recovery," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp.853–860, June 2012.

[80] F. Özyurt, E. Sert, E. Avci, and E. Dogantekin, "Brain tumor detection based on Convolutional Neural Network with neutrosophic expert maximum fuzzy sure entropy," *Measurement*, vol.147, pp.106–113, Dec. 2019.

[81] T. Perol, M. Gharbi, and M. Denolle, "Convolutional neural network for earthquake detection and location," *Science Advances*, vol.4, p.e1700578, Feb. 2018.

[82] S. Zhang, S. Zhang, C. Zhang, X. Wang, and Y. Shi, "Cucumber leaf disease identification with global pooling dilated convolutional neural network," *Computers and Electronics in Agriculture*, vol.162, pp.422–430, July 2019.

[83] X. Lu, H. Sun, and X. Zheng, "A Feature Aggregation Convolutional Neural Network for Remote Sensing Scene Classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol.57, pp.7894–7906, Oct. 2019.

[84] S.-W. Zhang, Y. Wang, X.-X. Zhang, and J.-Q. Wang, "Prediction of the RBP binding sites on lncRNAs using the high-order nucleotide encoding convolutional neural network," *Analytical Biochemistry*, vol.583, p.113364, Oct. 2019.

[85] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol.86, pp.2278–2324, Nov. 1998.

[86] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Deep Audio-visual Speech Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.40, no.7, pp.1–15, 2018.

[87] J. Gehring, M. Auli, D. Grangier, D. Yarats, and Y. N. Dauphin, "Convolutional sequence to sequence learning," in *Proceedings of the 34th International Conference on Machine Learning*, (Sydney, NSW, Australia), pp.1243–1252, Aug. 2017.

[88] T. Eady, “Tesla’s Deep Learning at Scale: Using Billions of Miles to Train Neural Networks,” Dec. 2019. [URL: https://towardsdatascience.com/teslas-deep-learning-at-scale-7eed85b235d3], [Accessed: 12/25/2019].

[89] D. Scherer, A. Müller, and S. Behnke, “Evaluation of Pooling Operations in Convolutional Architectures for Object Recognition,” in *Artificial Neural Networks – ICANN 2010*, pp.92–101, Springer Berlin Heidelberg, 2010.

[90] V. Nair and G. E. Hinton, “Rectified linear units improve restricted boltzmann machines,” in *Proceedings of the 27th International Conference on International Conference on Machine Learning*, pp.807–814, Omnipress, June 2010.

[91] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks,” in *Advances in Neural Information Processing Systems*, pp.1097–1105, Curran Associates, Inc., 2012.

[92] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, “ImageNet large scale visual recognition challenge,” *International Journal of Computer Vision (IJCV)*, vol.115, no.3, pp.211–252, 2015.

[93] Y. Cheng, “Mean shift, mode seeking, and clustering,” *IEEE transactions on pattern analysis and machine intelligence*, vol.17, no.8, pp.790–799, 1995.

[94] D. Comaniciu and P. Meer, “Mean shift: A robust approach toward feature space analysis,” *IEEE Transactions on pattern analysis and machine intelligence*, vol.24, no.5, pp.603–619, 2002.

[95] Y. Aliyari Ghassabeh, “A sufficient condition for the convergence of the mean shift al-

gorithm with Gaussian kernel," *Journal of Multivariate Analysis*, vol.135, pp.1–10, Mar. 2015.

[96] H. Freeman and R. Shapira, "Determining the minimum-area encasing rectangle for an arbitrary closed curve," *Communications of the ACM*, vol.18, pp.409–413, July 1975.

[97] E. Rashedi and H. Nezamabadi-pour, "A hierarchical algorithm for vehicle license plate localization," *Multimedia Tools and Applications*, vol.77, pp.2771–2790, Jan. 2018.

[98] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol.1, pp.500–511, IEEE, 2001.

[99] E. Rahtu, J. Heikkila, V. Ojansivu, and T. Ahonen, "Local phase quantization for blur-insensitive image analysis," *Image and Vision Computing*, vol.30, pp.501–512, Aug. 2012.

[100] G. Zhao, T. Ahonen, J. Matas, and M. Pietikainen, "Rotation-Invariant Image and Video Description With Local Binary Pattern Features," *IEEE Transactions on Image Processing*, vol.21, pp.1465–1477, Apr. 2012.

[101] M. Tkalcic and J. F. Tasic, "Colour spaces: perceptual, historical and applicational background," in *The IEEE Region 8 EUROCON 2003. Computer as a Tool.*, vol.1, pp.304–308, Sept. 2003.

[102] T. H. Vu, H. S. Mousavi, V. Monga, G. Rao, and U. A. Rao, "Histopathological image classification using discriminative feature-oriented dictionary learning," *IEEE transactions on medical imaging*, vol.35, no.3, pp.738–751, 2016.

[103] C. Liu, J. Yuen, A. Torralba, J. Sivic, and W. T. Freeman, "SIFT Flow: Dense Correspon-

dence across Different Scenes," in *Computer Vision – ECCV 2008*, pp.28–42, Springer, Oct. 2008.

[104] C. Wolf and J.-M. Jolion, "Object count/Area graphs for the evaluation of object detection and segmentation algorithms," *International Journal on Document Analysis and Recognition*, vol.8, no.4, pp.280–296, 2006.

[105] D. Karatzas, F. Shafait, S. Uchida, M. Iwamura, L. G. i Bigorda, S. R. Mestre, J. Mas, D. F. Mota, J. A. Almazan, and L. P. de las Heras, "ICDAR 2013 robust reading competition," in *2013 12th International Conference on Document Analysis and Recognition*, pp.1484–1493, IEEE, 2013.

[106] H. Amiri, H. Hojjat, and F. Oroumchian, "Investigation on a feasible corpus for Persian POS tagging," in *Proceedings of the 12th International CSI Computer Conference (CSICC)*, (Tehran), pp.20–26, 2007.

[107] R. Sibson, "SLINK: An optimally efficient algorithm for the single-link cluster method," *The Computer Journal*, vol.16, no.1, pp.30 – 34, 1973.

[108] N. Bodla, B. Singh, R. Chellappa, and L. S. Davis, "Soft-NMS — Improving Object Detection with One Line of Code," in *2017 IEEE International Conference on Computer Vision (ICCV)*, (Venice), pp.5562–5570, IEEE, Oct. 2017.

[109] T. Chen, M. Li, Y. Li, M. Lin, N. Wang, M. Wang, T. Xiao, B. Xu, C. Zhang, and Z. Zhang, "MXNet: A flexible and efficient machine learning library for heterogeneous distributed systems," 2015. arXiv: 1512.01274.

[110] J. Ding, N. Xue, Y. Long, G.-S. Xia, and Q. Lu, "Learning roi transformer for oriented object detection in aerial images," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[111] K. Fu, Z. Chang, Y. Zhang, G. Xu, K. Zhang, and X. Sun, “Rotation-aware and multi-scale convolutional neural network for object detection in remote sensing images,” *ISPRS Journal of Photogrammetry and Remote Sensing*, vol.161, pp.294 – 308, 2020.

Aabstract

This thesis aims at finding the text regions in Farsi natural scene text images. Text Localization is defined as locating regions in images that are considered as text by human beings. This problem is challenging and open because text can vary in size, font, and color, and the background can be non-uniform. As the Farsi text is intertwined and very similar to foliage, it is even more challenging to locate.

Four distinguished methods for localization of Farsi and Latin text are introduced in this thesis. Each of these methods addresses the weak points of the previous ones and improve upon them. The first method, Edge Color Signature, introduces colored edges for locating candidate text regions. First, the Mean-Shift clustering algorithm is employed by this method for assigning colors to edge pixels. Consequently, these colored edges are utilized to produce candidate text regions. Finally, a cascade classifier is used to classify the candidate regions. This classifier employs novel features based on edge pixels and dictionary learning methods to detect text regions. The second method, Edge Color Transform, improves the edge extraction method of Edge Color Signature. Also, it introduces a novel, fast and accurate operator for producing colored edges.

The third method, Deep Color Transform, improves the second method to locate text with non-uniform backgrounds. Also, a convolutional neural network with novel architecture is introduced to classify the candidate regions. Finally, the fourth method, called Meta-Heuristic Net, uses the convolutional network of the third method; The network's output is employed in a meta-heuristic search method to locate text regions. In this search method, search windows are distributed over the input image and gradually converge toward the text regions.

The proposed methods are examined and compared with each other and several state-of-the-art methods on two datasets: Farset for Farsi text and ICDAR2013 for English text. The results show that the Deep Color Transform with the f-measure value of 86.64 for English and 91.58 for Farsi images outperforms all of the other methods.

Keywords: Persian natural scene text localization; natural scene; Edge color transform; Edge growing; Deep Learning; Meta-heuristic search method

Shahrood University of Technology

Faculty of Computer Engineering

Ph.D. Thesis in Artificial Intelligence Engineering

Persian Natural Scene Text Localization Using Meta-Heuristic Search Method

By: Jalil Ghavidel Neycharan

Supervisor

Dr. Alireza Ahmadyfard

Advisor

Dr. Morteza Zahedi

February 2020