مجله ماشین‌بینایی و پردازش‌تصویر
سال پنجم، شماره اول، تابستان ۱۳۹۷

بخش‌بندی معنایی نظارتی ضعیف با استفاده همزمان از اطلاعات سطح شی و سطح متن

پروین رزاقی*

چکیده

در این مقاله، روش جدیدی برای بخش‌بندی معنایی تصاویر در حضور داده‌های آموزشی نظارتی ضعیف ارائه می‌گردد. هدف اصلی در بخش‌بندی معنایی اختصاص برچسب به تمامی پیکسل‌های تصویر است. در داده‌های آموزشی نظارتی ضعیف، تنها برچسب‌های معنایی موجود در تصویر مشخص می‌گردد و مکان آن‌ها در تصویر مشخص نمی‌گردد. نوآوری روش پیشنهادی، استفاده همزمان از اطلاعات سطح شی و سطح متن در تعیین برچسب‌های معنایی در تصویر می‌باشد. در روش پیشنهادی، نواحی تصاویری که دارای مجموعه برچسب‌های یکسانی می‌باشند، با یکدیگر ترکیب می‌گردند به گونه‌ای که در تصاویری که دارای برچسب‌های مشترک هستند، نحوه ظهور یکسان داشته و موقعیت مکانی آن‌ها نسبت به دیگر برچسب‌های معنایی موجود در تصویر نیز یکسان باشد. همچنین برای بهینه کردن تابع هزینه‌ی پیشنهادی، یک الگوریتم تکرار شونده ارائه شده است که در آن در ابتدا تمامی پیکسل‌های مجموعه تصاویر، به صورت اولیه برچسب گذاری می‌گردد. سپس مدل ظهور هر برچسب معنایی و مدل متن آن با استفاده از ماشین بردار پشتیبان آموزش می‌بینند. در قدم بعد، برچسب پیکسل‌ها به گونه‌ای به روزرسانی می‌گردد که در مجموعه تصاویری که دارای برچسب‌های یکسانی می‌باشند، اطلاعات سطح شی و سطح متن مشابه باشند. به روزرسانی برچسب‌ها تا زمانی ادامه می‌یابد که در دو دور متوالی، برچسب پیکسل‌ها تغییر نیابد. برای ارزیابی کارایی روش پیشنهادی از مجموعه داده‌ی MSRC استفاده شده است. روش پیشنهادی بر روی مجموعه داده‌ی MSRC، دقت میانگین نرخ شناسایی گروهی ۷۲% را به دست آورده است که در مقایسه با دیگر روش‌های قابل مقایسه و موفق پیشین ۱% افزایش دقت داشته است.

کلید واژه ها

بخش‌بندی معنایی نظارتی ضعیف، اطلاعات سطح شی، اطلاعات سطح متن، الگوریتم بسط حرکت.

۱ مقدمه

یکی از اهداف بلند مدت در حوزه‌ی هوش مصنوعی، طراحی رباتی است که قادر باشد با محیط اطراف خود تعامل داشته و ارتباط برقرار نماید. بدین منظور ربات از ادراک متفاوتی همانند بینایی، شنوایی، بویایی و غیره استفاده می‌نماید. یکی از مهمترین ادراک‌ها برای تعامل با محیط، بینایی می‌باشد. از اینرو حوزه‌ی بینایی کامپیوتر یکی از مهمترین حوزه‌های تحقیقاتی در هوش مصنوعی است. یکی از مهمترین قدم‌ها برای حرکت به سمت درک محیط از طریق بینایی، بخش‌بندی معنایی صحنه‌ای است که مشاهده می‌گردد. بخش‌بندی معنایی یکی از حوزه‌های مهم تحقیق در بینایی کامپیوتر است. نتایج بخش‌بندی معنایی کاربردهای فراوانی دارد که از جمله‌ی آن‌ها می‌توان به موتور جستجوی تصاویر، راننده خودکار و تعامل انسان و ماشین نام برد. هدف اصلی در بخش‌بندی معنایی، اختصاص برچسب به تمامی پیکسل‌های موجود در تصویر می‌باشد. برچسب‌ها بیانگر اطلاعات معنایی (همانند آسمان، جاده، سبزه‌زار، شی پس زمینه و غیره) می‌باشند. روش‌های بخش‌بندی معنایی براساس نوع اطلاعات

این مقاله در آبان‌ماه ۱۳۹۵ دریافت، در فروردین‌ماه ۱۳۹۶ بازنگری و در خردادماه همان سال پذیرفته شد.

* دانشکده علوم رایانه و فناوری اطلاعات، دانشگاه تحصیلات تکمیلی علوم پایه زنجان
رایانامه: p.razaghi@iasbs.ac.ir

موجود در نمونه‌های آموزشی به دو دسته روش‌های نظارتی[1] و روش‌های نظارتی ضعیف[2] تقسیم می‌گردند. در روش‌های نظارتی، تمامی پیکسل‌های تصاویر نمونه‌های آموزشی توسط کاربر انسانی برچسب خورده است. در حالی که در روش‌های نظارتی ضعیف به هر تصویر چندین برچسب متنی[3] اختصاص داده شده است که بیانگر دسته‌های معنایی موجود در تصویر است. روش‌های نظارتی و نظارتی ضعیف نیز خود براساس نحوه‌ی حل مسئله به دو دسته پارامتریک و غیرپارامتریک تقسیم می‌گردند [۱] که در ادامه هر کدام از روش‌های پارامتریک و غیرپارامتریک تعریف گردیده و روش‌های موجود در هر دسته مورد بررسی و تحلیل قرار می‌گیرند. روش‌های غیرپارامتریک در سال‌های اخیر بیشتر مورد توجه قرار گرفته است. زیرا این روش‌ها وابسته به مجموعه داده و مدل‌های یادگیری خاصی نمی‌باشند و با افزودن نمونه‌های آموزشی جدید و یا دسته معنایی جدید نیازی به تغییر مدل یاد گرفته شده نداریم.

در روش‌های پارامتریک، مدل یادگیری، تابعی با فرم مشخص دارد که دارای پارامترهایی نیز می‌باشد که پارامترها در طی مرحله‌ی آموزش به دست می‌آیند [۲]. سپس در مرحله‌ی آزمایشی، از مدل یاد گرفته شده استفاده می‌گردد. از اینرو روش‌های پارامتریک وابسته به نمونه‌های آموزشی می‌باشند. پارامترهای مدل هنگامی که نمونه جدیدی به سیستم اضافه می‌گردد بایستی به روز رسانی شود. همچنین، در اکثر موارد در صورت مواجهه با یک مجموعه داده جدید، مدل‌ها بایستی دوباره آموزش داده شوند. تاکنون روش‌های پارامتریک زیادی معرفی گردیده است. اکثر روش‌های بخش‌بندی معنایی از میدان تصادفی شرطی استفاده می‌نمایند که بر روی نواحی تعریف می‌گردد و در حالت استاندارد دارای دو عبارت می‌باشد. عبارت داده که اطلاعات ظاهری هر ناحیه را در نظر می‌گیرد و عبارت همواری که نواحی همسایه و مشابه را ترغیب می‌نماید که دارای برچسب‌های یکسانی باشند. در میدان تصادفی شرطی استاندارد، وابستگی‌های سطح بالا میان رأس‌های گراف (رأس‌های گراف متناظر با پیکسل و یا نواحی تصویر می باشد) در نظر گرفته نمی‌شود. به همین دلیل، در سال‌های اخیر، اکثریت روش‌های ارائه شده سعی در استفاده از اطلاعات سطح بالا در مسئله دارند.

گلد در رساله دکتری خود [۳] روش جدیدی برای انتخاب نواحی مناسب برای تصویر ارائه داده است. او در ابتدا با استفاده از الگوریتم‌های بخش‌بندی تصویر (همانند الگوریتم‌های تعیین فراپیکسل)، مجموعه ناحیه‌های متفاوتی برای تصویر تعریف می‌نمایند که هر کدام از این مجموعه‌ها نمایش کاملی از تصویر را فراهم می‌نمایند. سپس یکی از این مجموعه ناحیه‌ها به عنوان بخش‌بندی اولیه تصویر انتخاب می‌شود و مقدار تابع انرژی برای این مجموعه به دست می‌آید. در قدم بعد، برای اصلاح ناحیه‌های موجود و انتخاب ناحیه‌های مناسب از الگوریتم استنتاج انتخاب حرکت[4] استفاده می‌نماید. در این روش از استنتاج، مجموعه‌ای از تغییرات در ناحیه‌های تصویر در نظر گرفته می‌شود، سپس این تغییرات رتبه‌بندی شده و تغییرات بر حسب رتبه اعمال شده و مقدار تابع انرژی نیز پس از تغییر محاسبه می‌شود. در صورتی که مقدار تابع انرژی کاهش یابد، تغییر اعمال شده بر روی بخش‌بندی تصویر اعمال می‌گردد و در غیر اینصورت تغییر مورد نظر پذیرفته نمی‌شود. این رویه تا زمانی ادامه می‌یابد که هیچ کدام از تغییرات موجب کاهش تابع انرژی نشوند.

در [۴] روشی ارائه شده است که مجموعه‌ای از بهترین نواحی تصویر که قادر به ارائه نمایش کاملی از مسئله هستند از طریق کمینه نمودن تابع انرژی به دست می‌آید. روش ارائه شده شامل یک شبکه دو لایه است. در لایه اول، به هر پیکسل برچسبی اختصاص داده می‌شود که بیانگر این است که متعلق به چه ناحیه‌ای می‌باشد. سپس در لایه دوم، به نواحی انتخاب شده، برچسب یکی از اشیای موجود در صحنه تعلق می‌پذیرد. با توجه به اینکه در لایه اول تعداد نواحی ممکن برای هر پیکسل نمایی می‌باشد، استنتاج عملی نخواهد بود. به همین دلیل، در ابتدا فرهنگ لغات بزرگی از نواحی در نظر گرفته می‌شود. در نتیجه، در لایه اول انتخاب نواحی از روی این فرهنگ لغات صورت می‌پذیرد. نحوه انتخاب نواحی در فرهنگ لغات محدودیت خاصی ندارد، تنها بایستی مجموعه‌ای از نواحی در فرهنگ لغات موجود باشد که قادر به ارائه‌ی توصیف کاملی از تصویر باشند.

لدیکی [۵] در رساله دکتری خود سوال مهم "بهترین بخش‌بندی تصویر کدام است و چگونه می‌توان به آن دست یافت؟" را مطرح نموده است. همچنین عنوان کرده است که می‌توان ادعا کرد که پیدا کردن یک بخش‌بندی ایده‌آل برای تصویر ممکن نبوده و بخش‌بندی کردن تصاویر متفاوت نیازمند الگوریتم‌های متفاوتی می‌باشد. لدیکی [۵] برای حل این مشکل روشی سلسله مراتبی مبتنی بر میدان تصادفی شرطی (CRF) ارائه داده است. آن‌ها بخش‌بندی‌های متفاوت را به گونه‌ای در کنار یکدیگر قرار می‌دهند که دیگر نیازی به این تصمیم نداریم که کدام از یک از این بخش‌بندی‌ها مناسب می‌باشد. بایستی دقت کرد که در روش مذکور به طور صریح نواحی مطلوب از لایه‌های مختلف انتخاب نمی‌شوند. بلکه برچسب گذاری در پایین‌ترین سطح (سطح پیکسل) در صورتی قابل قبول است که برچسب گذاری‌های حاصل در لایه‌های بالاتر، با توجه به برچسب گذاری پایین‌ترین لایه، خطای کمتری (خطای کم با پایین بودن مقدار تابع انرژی هم‌ارز است) را تولید نماید. همچنین تابع انرژی بر روی تمامی لایه‌ها تعریف می‌شود. در این حالت تابع انرژی در هر لایه دارای سه عبارت پتانسیل می‌باشد. عبارت اول تضمین می‌نماید که پیکسل‌هایی که در درون ناحیه یکسان قرار دارند دارای برچسب یکسانی باشند. عبارت دوم نیز نواحی همسایه در لایه‌های مجاور را ترغیب می‌نماید که دارای

[1] Supervised
[2] Weakly supervised
[3] Image tags
[4] Move making

برچسب یکسانی باشند. عبارت سوم کمترین میزان انرژی حاصل از برچسب گذاری نواحی موجود در لایه بالاتر را، با استفاده از برچسب‌های لایه فعلی محاسبه می‌نماید.

برخی از روش‌ها از نتایج حاصل از تشخیص‌دهنده‌های شی استفاده می‌نمایند. دانش حاصل از روش‌های دیگر (همانند شناسایی اشیا) در برچسب گذاری تصویر، در عبارت پتانسیل محلی مدل می‌گردد. در بسیاری از روش‌ها فرض می‌گردد که نتایج حاصل از الگوریتم‌های تشخیص شی صحیح بوده و هیچ خطایی در آن‌ها وجود ندارد. در نتیجه در صورت بروز خطا، نمی توان از رخداد آن جلوگیری کرد. در [۵] روشی ارائه شده است که بر این مشکل فائق آمده است. در این روش، به نتایج حاصل از الگوریتم‌های تشخیص شی متغیری دودویی نسبت داده می‌شود. در صورتی که متغیر مقدار یک بگیرد، نتیجه حاصل از شناسایی شی مورد قبول واقع می‌شود. در این حالت عبارت داده دیگری نیز به تابع انرژی اضافه می‌گردد، در حالی که تابع انرژی مبتنی بر پیکسل و یا سوپرپیکسل به قوت خود باقی است. به عبارت دیگر، از شناسایی کننده‌های اشیا به عنوان یک اطلاعات اضافی استفاده می‌شود. یااو و همکارانش در [۶] برای بخش‌بندی اولیه تصویر از الگوریتم ارائه شده در [۷] استفاده می‌نمایند. برای انجام این کار از یک شبکه دو لایه از ناحیه‌ها استفاده می‌نمایند. در لایه اول آستانه الگوریتم آب‌پخشان[1] در روش [۶] کوچک در نظر گرفته می‌شود، در نتیجه تعداد ناحیه‌های ایجاد شده زیاد است. در حالی که در لایه دوم آستانه بزرگتر در نظر گرفته می‌شود، در نتیجه تعداد ناحیه‌ها در لایه دوم کاهش می‌یابد. ناحیه‌های لایه دوم فراناحیه[2] نامیده می‌شود. در واقع روش ارائه شده در [۶] با استفاده از این کار سعی در معنادارتر کردن نواحی ایجاد شده دارد. همچنین در این روش، از شناسایی کننده شی نیز استفاده می‌گردد. همانند روش قبل به هر خروجی الگوریتم شناسایی شی یک متغیر دودویی نسبت داده می‌شود که بیانگر صحیح و یا غیر صحیح بودن نتیجه حاصل است. در این روش، علاوه بر شناسایی شی، نوع صحنه نیز به عنوان یک معیار اضافی دیگر استفاده می‌شود. به همین منظور، ابتدا نوع صحنه با استفاده از الگوریتم‌های رده‌بندی صحنه به دست می‌آید. همچنین، از عبارت‌های اضافی دیگری که بیانگر سازگاری میان نوع صحنه و اشیا موجود در آن می‌باشد، نیز استفاده می‌گردد.

در [۸] برای استفاده از اطلاعات متن[3] مجموعه رئوس گراف متناظر با اشیا موجود در تصویر و مجموعه‌ی یال‌ها متناظر با ارتباط میان اشیا (اطلاعات هم‌رخدادی اشیا) می‌باشد. آن‌ها در ابتدا تصویر را به یکسری از نواحی تجزیه می‌کنند. اما برای استدلال برچسب‌های پیکسل‌های تصویر از این نواحی استفاده نمی‌نمایند. بلکه استدلال در سطح شی صورت می‌پذیرد که هر شی، مجموعه‌ای از نواحی در نظر گرفته می‌شود. حال، برای استدلال از مدل ساختار مصور استفاده می‌نماید. در مدل ساختار مصور، پارامترهای مدل براساس یک ساختار مشخص در گراف به دست می‌آیند. در حالی که در مسئله برچسب گذاری، ساختار میان پاره‌ها (در اینجا، پاره‌ها متناظر با اشیا می‌باشند) دارای فرم مشخصی (همانند درخت، فرم ستاره‌ای و یا غیره) نمی‌باشند. به همین دلیل پارامترهای مدل به گونه‌ای تغییر داده شده‌اند که مستقل از ساختار میان پاره‌ها باشند. پارامترهای مدل در حالت عبارت‌های یکانی و دوتایی به دست می‌آیند. در عبارت‌های یکانی، پارامترهای نحوه ظهور[4]، شکل و موقعیت اشیا به دست می‌آیند. در عبارت‌های دوتایی نیز، پارامترهای فاصله و زاویه میان پاره‌ها (اشیا) به دست می‌آید. در قدم استدلال، ابتدا بایستی ساختار میان پاره‌ها (اشیا) استخراج گردد. کورسو [۸] فرض می‌کند که این ساختار را برای هر تصویر آزمایشی در اختیار دارد. سپس براساس ساختار تعیین شده، برچسب گذاری پیکسل‌ها صورت می‌پذیرد. یکی از مهمترین معایب این روش، وابستگی آن به کاربر می‌باشد. زیرا برای هر تصویر آزمایشی بایستی گراف ساختار میان پاره‌ها (اشیا) توسط کاربر فراهم گردد.

بخش‌بندی معنایی تصویر در حضور داده‌های نظارتی ضعیف دارای اهمیت بسیار می‌باشد. زیرا فراهم کردن داده‌های نظارتی کامل در فاز آموزش بسیار هزینه بر است. همچنین با توجه به اینکه اکثریت تصاویر موجود در اینترنت دارای برچسب‌های متنی هستند که بیانگر اشیای موجود در تصویر است، با ارائه راه‌حل مناسب برای بخش بندی معنایی با استفاده از داده‌های نظارتی ضعیف می‌توان از این حجم عظیم از داده‌ها استفاده نمود.

تاکنون کارهای زیادی در حوزه بخش‌بندی معنایی با استفاده از داده‌های نظارتی ضعیف انجام شده است که با بررسی تمامی کارهایی که تاکنون صورت گرفته است می‌توان به این نتیجه رسید که تاکنون هیچ کدام از کارهای انجام شده قادر نبوده‌اند از اطلاعات سطح شی و سطح متن[5] به طور موثری در فاز استنتاج استفاده نمایند. وژنونت و همکارانش [۹] برای بخش‌بندی معنایی با استفاده از داده‌های نظارتی ضعیف، یک مدل گرافیکی جدیدی پیشنهاد داده‌اند که در آن علاوه بر ارتباطات ناحیه‌های مجاور موجود در یک تصویر، نواحی مشابه میان مجموعه تصاویر که دارای برچسب یکسان می‌باشند نیز دارای ارتباط می‌باشند. مدل گرافیکی پیشنهاد شده قادر است اطلاعات سطح ناحیه را میان تصاویر مختلف در نظر بگیرد اما با توجه به اینکه هر ناحیه در تصویر بخش کوچکی از تصویر در نظر گرفته شده که پیکسل‌های موجود در آن دارای خاصیت مشترکی می‌باشند، در نتیجه روش مذکور قادر به در نظر گرفتن اطلاعات سطح شی و سطح متن نمی‌باشد. آن‌ها برای غلبه بر مشکل مذکور در توصیفگرهای

[1] Watershed
[2] Super region
[3] Contextual Information
[4] Appearance
[5] Context

نواحی، با استفاده از روش [۱۰] میزان متعلق بودن ناحیه به یک شی را نیز به عنوان یکی از مولفه‌های توصیفگر در نظر گرفته اند. با توجه به اینکه احتمال خطا در خروجی حاصل از روش [۱۰] وجود دارد، در صورت رخداد خطا، خطا در کل روش انتشار پیدا خواهد کرد.

لیو و همکارانش [۱۱] برای حل مسئله، معیارهای مورد نظر برای حل مسئله را در فرم تابع انرژی در نظر گرفته، سپس با کمینه نمودن تابع انرژی، بخش‌بندی معنایی برای هر تصویر به دست می‌آید. آن‌ها اطلاعات مشابه میان نواحی تصاویری که دارای برچسب مشترک می‌باشند را در عبارت پیشنهادی خود در نظر گرفته است. بدین منظور از تابع انرژی روش خوشه‌بندی طیفی استفاده می‌نماید. برای در نظر گرفتن ویژگی‌های مهم هر گروه (خوشه) در تصمیم‌گیری، از خوشه‌بندی متمایزکننده استفاده می‌نماید. سپس برخی از محدودیت‌های بخش‌بندی معنایی در نظر گرفته می‌شود. اعم از اینکه هر ناحیه از تصویر تنها می‌تواند یک برچسب معنایی به خود اختصاص دهد و یا مجموعه برچسب‌های اختصاصی به هر تصویر بایستی با تعداد برچسب‌های متنی تصویر برابر باشد. تابع انرژی حاصل یک تابع انرژی محدب نمی‌باشد. بدین منظور برای بهینه سازی کردن تابع انرژی از روش CCCP استفاده می‌نماید.

کیم و همکارانش [۱۲] سعی در بدست آوردن یک شی متمایز میان مجموعه‌ای از تصاویر هستند که در آن مجموعه از تصاویر تنها گروه‌های شی موجود در تصویر ذکر شده است. آن‌ها برای حل مسئله مذکور، هر تصویر را با استفاده از معادله گرما مدل نموده‌اند و سپس با بیشینه نمودن دمای کلی تصویر سعی در یافتن مکان شی متمایز در تصویر دارند. ژولین [۱۳] روش جدیدی برای بخش‌بندی همزمان ارائه داده است که در بخش‌بندی همزمان قادر است وجود چندین شی متمایز در تصویر را مدیریت نماید. زینگ و همکارانش [۱۴] روش ارائه داده شده در [۹] را بسط داده و اطلاعات متن برای هر سوپرپیکسل‌ها را در آن استفاده نموده‌اند. بدین منظور، برای هر فراپیکسل، اطلاعات برچسب‌های معنایی فراپیکسل‌های همسایه‌ی آن نیز در نظر گرفته می‌شود. در [۱۵]، اطلاعات نظارتی ضعیف به سه دسته تقسیم شده است: ۱) اطلاعات متنی که شامل اشیای موجود در تصویر است ۲) مستطیل‌های محاط دور شی و ۳) برچسب‌گذاری نسبی پیکسل‌ها. آن‌ها یک روش یکپارچه‌ای ارائه داده‌اند که در صورت مواجهه با هر نوع اطلاعات در دسترس قادر است که برچسب‌گذاری پیکسل‌ها را انجام دهد. آن‌ها از خوشه‌بندی حاشیه بیشینه استفاده نمودند و نوع اطلاعات نظارتی ضعیف را با استفاده از یکسری محدودیت نمایش دادند. نیو و همکارانش [۱۶] روشی برای بخش‌بندی معنایی نظارتی ضعیف در حضور برچسب‌های متنی خدشه‌دار ارائه داده‌اند. آن‌ها ابتدا تصویر را به مجموعه‌ای از ناحیه‌ها با استفاده از برچسب‌های متنی خدشه‌دار تقسیم می‌نمایند. سپس برای در نظر گرفتن اطلاعات خدشه‌دار، ماتریس شامل برچسب‌های اولیه ناحیه‌ها به ماتریس مرتبه پایین تجزیه می‌گردد. برای انجام این کار از عامل بندی ماتریس بهره گرفته شده است و برای نیل به پاسخ بهتر، عبارت منظم‌سازی لاپلاسین به عبارت عامل بندی ماتریس اضافه شده است. سپس عبارت حاصل با استفاده از الگوریتم انتشار برچسب حل شده است. صالح و همکارانش [۱۷]، در ابتدا با استفاده از خروجی لایه‌های پیچش شبکه عصبی پیچش و میدان تصادفی شرطی، تصویر را به دو بخش پیش زمینه و پس زمینه تقسیم می‌نماید. سپس تابع هزینه جدیدی برای بخش‌بندی معنایی نظارتی ضعیف ارائه داده‌اند که اطلاعات پیش‌زمینه و پس‌زمینه استخراج شده در آن دخالت داده شده است. همچنین با توجه اینکه اطلاعات پیش‌زمینه و پس زمینه بدست آمده دارای خطا است، به همین دلیل چندین نقاب برای پیش‌زمینه و پس‌زمینه ایجاد می‌گردد تا کاربر بهترین آن‌ها را انتخاب نماید.

یکی از مهمترین چالش‌های روش‌های بخش‌بندی نظارتی ضعیف پیشین، استفاده نکردن از اطلاعات سطح بالا به صورت صریح است. به عبارتی دیگر، آن‌ها در تعیین برچسب معنایی برای ناحیه‌های محلی تنها از نحوه‌ی ظهور خود ناحیه و ناحیه‌های مشابه در تصاویر دیگر استفاده نموده‌اند. در حالی که اطلاعات محلی به تنهایی برای تعیین برچسب معنایی کافی نبوده و برای تعیین دقیق برچسب نیازمند اطلاعات سطح شی و سطح متن می‌باشیم. در این مقاله روش جدیدی برای بخش‌بندی نظارتی ضعیف ارائه شده است که اطلاعات سطح شی و سطح متن به صورت صریح در آن استفاده داده شده است. بدین منظور در تابع هزینه‌ی پیشنهادی، برچسب به نواحی تصویر به گونه‌ای اختصاص می‌یابد که توصیفگر ناحیه‌های همبندی از تصویر که به یک برچسب اختصاص داده شده‌اند به مدل سطح شی برچسب متناظر مشابه باشد و چیدمان برچسب‌های معنایی به مدل چیدمان آموزش دیده مشابه باشد. مدل سطح شی، مدل ظهور برچسب معنایی را در سطح شی در نظر می‌گیرد. مدل سطح متن، مدل آموزش دیده‌ی چیدمان برچسب‌های معنایی در سطح تصویر است.

مهمترین نوآوری‌های روش پیشنهادی نسبت به روش‌های موفق پیشین عبارت است از ۱) استفاده از اطلاعات سطح شی ۲) استفاده از اطلاعات سطح متن ۳) انتقال دانش میان مجموعه تصاویری که دارای برچسب‌های معنایی یکسانی می‌باشند.

برای ارزشیابی روش‌های بخش‌بندی معنایی از مجموعه داده‌های متفاوتی استفاده می‌گردد که از جمله‌ی آن‌ها می‌توان به MSRC، LMO و پاسکال اشاره نمود. در این مقاله برای ارزیابی روش پیشنهادی از مجموعه داده‌ی MSRC بهره گرفته شده است. مجموعه داده‌ی MSRC هم شامل اشیا ساختار یافته همانند انسان، خودرو، دوچرخه بوده و هم شامل اشیا غیرساختاریافته

همانند سبزه، آب و آسمان می‌باشد. در حالیکه مجموعه داده‌ی LMO شامل تصاویر صحنه‌های بیرونی همانند کوه، جنگل، جاده و ساحل می‌باشد. مجموعه داده‌ی پاسکال شامل اشیا ساختاریافته می باشد. یکی از مشکلات مجموعه داده‌ی پاسکال، ناقص بودن اطلاعات برچسب‌های متنی متناظر با هر تصویر است. باتوجه به اینکه روش پیشنهادی اطلاعات سطح شی و سطح متن را به طور همزمان در نظر می‌گیرد، در این مقاله برای ارزیابی روش پیشنهادی از مجموعه داده‌ی MSRC استفاده شده است.

در این مقاله، در بخش ۲، مسئله به صورت دقیق تعریف می‌گردد و حل مسئله به صورت یک تابع مدل می‌گردد. جهت بهینه سازی تابع، یک روش تکراری ارائه می‌گردد. برچسب گذاری اولیه تصاویر و یادگیری مدل اولیه برای نحوه ظهور و اطلاعات متن هر گروه معنایی در بخش ۲-۱ بیان می‌گردد. سپس در بخش ۲-۲، روش پیشنهادی برای بهینه سازی تابع ارائه می‌گردد. در بخش ۳، روش پیشنهادی بر روی مجموعه داده استاندارد MSRC اعمال می‌گردد و نتایج گزارش می‌گردد. در نهایت، نتیجه گیری در بخش ۴ بیان می‌گردد.

۲ روش پیشنهادی

فرض کنید که $\{I_j\}_{j=1}^{N}$ بیانگر مجموعه تصاویری است که هر تصویر دارای چندین برچسب متنی است که برچسب‌های متنی مذکور، اشیا موجود در تصویر را نمایش می دهد و با استفاده از $\{t_j\}_{j=1}^{N}$ نمایش داده می‌شود. برچسب‌های متنی هر تصویر $t_j \in P(\{1,2,...,L\})$ است که در آن L تعداد برچسب‌های متنی معنایی موجود در کل مجموعه تصاویر است و $P(.)$ بیانگر مجموعه توانی است. هر تصویر به مجموعه ای از نواحی تقسیم می‌گردد که مجموعه نواحی تصویر jام با استفاده از $\{r_k^j\}_{k=1}^{s_j}$ نمایش داده می‌شود. پارامتر s_j بیانگر تعداد نواحی موجود در تصویر jام می‌باشد. در روش پیشنهادی برای تقسیم هر تصویر به مجموعه‌ای از نواحی از روش [۷] استفاده شده است. زیرا در روش مذکور نواحی ایجاد شده تا حد زیادی نسبت به دیگر روش-های موجود معنادارتر می‌باشد و این امر باعث می‌گردد که تعداد نواحی ایجاد شده در مقایسه با دیگر روش‌ها بسیار کمتر باشد. تعداد نواحی موجود در N تصویر موجود با استفاده از $S = s_1 + s_2 + ... + s_N$ نمایش داده می‌شود. هدف اصلی در بخش‌بندی معنایی نظارتی ضعیف، اختصاص برچسب معنایی به تمامی نواحی تصویر است. به عبارت دیگر هدف تعیین $\{\{l_k^j\}_{k=1}^{s_j}\}_{j=1}^{N}$ است که در آن l_k^j بیانگر برچسب معنایی ناحیه kام از تصویر jام است. در روش پیشنهادی برای رسیدن به هدف مذکور چندین هدف در نظر گرفته شده است که در ادامه عنوان می‌گردد:

۱. در صورت وجود گروه معنایی مشترک میان چندین تصویر، بایستی ویژگی‌های ظاهری ناحیه‌های مرتبط به گروه معنایی یکسان به یکدیگر مشابه باشد. با استفاده از این هدف در روش پیشنهادی، اطلاعات سطح شی در نظر گرفته می‌شود.

۲. تصاویری که دارای مجموعه برچسب‌های معنایی یکسانی هستند بایستی برچسب‌های معنایی پیکسل‌ها در تصاویر مختلف دارای چیدمان مکانی یکسانی باشند. این هدف اطلاعات سطح متن را در نظر می‌گیرد.

در روش پیشنهادی برای در نظر گرفتن اطلاعات هدف اول از عبارت زیر استفاده شده است:

$$\sum_{j=1}^{N}\sum_{l=1}^{L}\left\|d_{i_j(l)} - m_l\right\|^2 \qquad (۱)$$

که $i_j(l)$ بیانگر اندیس ناحیه‌هایی در تصویر j است که دارای برچسب l است و $d_{i_j(l)}$ بیانگر توصیفگر ناحیه‌های مورد نظر است. همچنین m_l نیز بیانگر مدل آموزش دیده برچسب معنایی l است. در عبارت مذکور، $i_j(l)$ و m_l جز مجهول‌های مسئله هستند که در طی مرحله‌ی آموزش به دست خواهد آمد. همچنین برای هدف دوم و در نظر گرفتن اطلاعات متن در تعیین برچسب برای هر پیکسل، عبارت زیر پیشنهاد شده است:

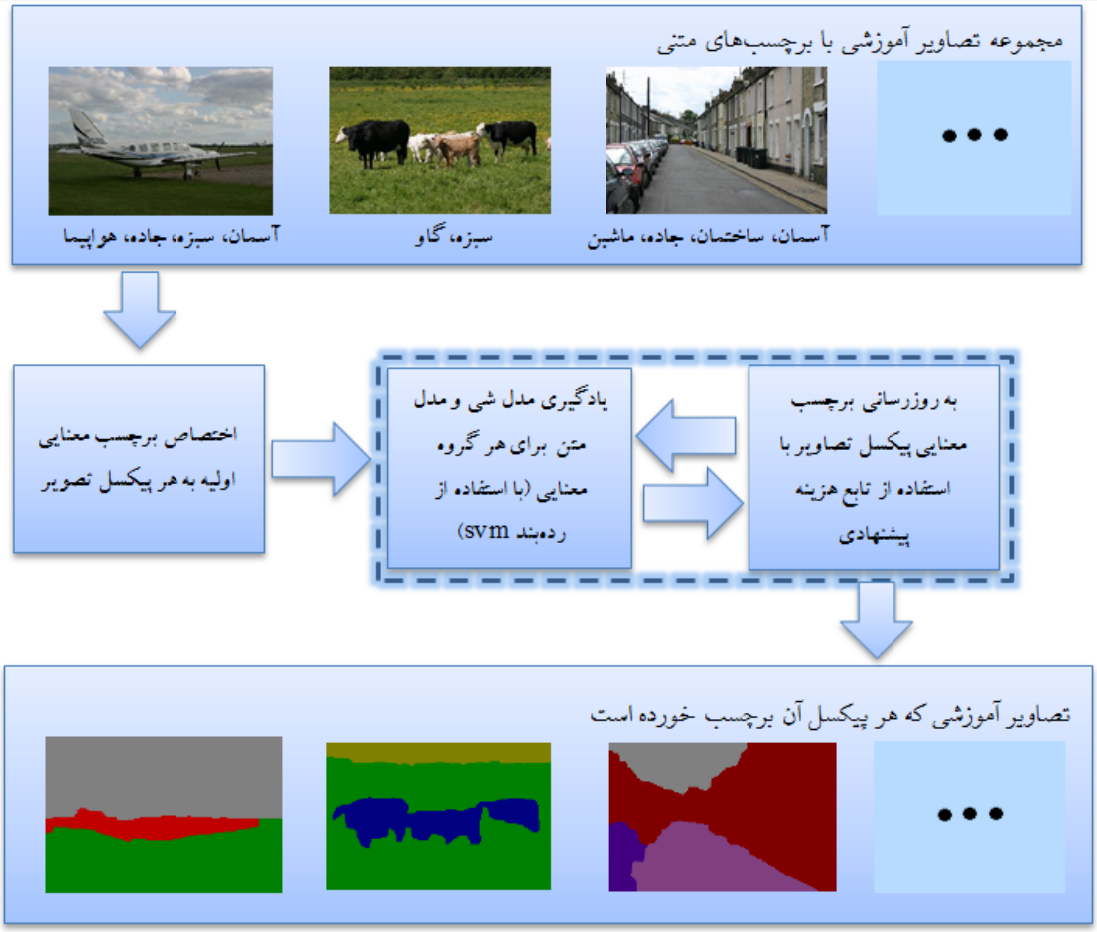
$$\sum_{j=1}^{N}\sum_{j'=j+1}^{N}\left\|C_j - C_{j'}\right\|^2 \qquad (۲)$$

که C_j بیانگر توصیفگر متن[1] تصویر است که اطلاعات چیدمانی برچسب‌های معنایی تصویر را در نظر می‌گیرد. هدف اصلی، به دست آوردن برچسب‌های پیکسل‌های تصویر است به گونه‌ای که تابع زیر کمینه گردد:

$$\{\{l_k^j\}_{k=1}^{n_j}\}_{j=1}^{N} = \arg\min \sum_{j=1}^{N}\sum_{j'=j+1}^{N}\sum_{l=1}^{L}\left\|d_{i_j(l)} - m_l\right\|^2 + \sum_{j=1}^{N}\sum_{j'=j+1}^{N}\left\|C_j - C_{j'}\right\|^2 \qquad (۳)$$

رابطه‌ی ۳ به گونه‌ای طراحی شده است که در برچسب گذاری مطلوب، نحوه‌ی ظهور گروهی از نواحی همسایه با برچسب معنایی یکسان به نحوه‌ی ظهور برچسب معنایی مذکور مشابه باشد (اطلاعات سطح شی). همچنین نحوه‌ی چیدمان برچسب‌های معنایی نیز باید با مدل‌های آموزش دیده مشابهت داشته باشد (اطلاعات سطح متن). از این رو، روش پیشنهادی قادر است اطلاعات سطح شی و اطلاعات سطح متن را به طور همزمان در نظر بگیرد. شکل ۱، شمای کلی روش پیشنهادی را نمایش می‌دهد.

[1] Context descriptor



شکل۱ شمای کلی روش پیشنهادی

با توجه به اینکه برچسب‌های معنایی هر تصویر مجهول می‌باشد و برای محاسبه توصیفگرهای سطح شی و سطح متن به اطلاعاتی نیازمندیم که وابسته به تعیین برچسب‌ها می‌باشد، نتیجه گرفته می‌شود که رابطه‌ی ۳، تابع غیر محدب است. در روش پیشنهادی برای بهینه نمودن آن از ایده مبتنی بر حرکت استفاده شده است که یک الگوریتم تکرار شونده است. بدین منظور فرض می‌گردد که یک برچسب گذاری اولیه از تصویر در اختیار داریم. سپس برای هر برچسب معنایی α الگوریتم بسط حرکت صورت می‌پذیرد. این کار تا زمانی انجام می‌گردد که برچسب پیکسل‌های تصویر تغییری نیابد. در الگوریتم بسط حرکت برای برچسب α، برای تعیین برچسب جدید برای پیکسل‌ها، در تصویر هر پیکسل تنها می‌تواند برچسب α را بپذیرد و یا اینکه همان برچسب قبلی باقی بماند. در روش پیشنهادی برای محدود کردن هر تصویر به مجموعه برچسب در هر مرحله از مرحله‌ی آموزش، مجموعه تصاویری که دارای مجموعه برچسب یکسانی هستند به الگوریتم داده می‌شود. در ادامه هر کدام از مراحل با جزئیات بیشتر شرح داده می‌شود.

۲-۱ برچسب گذاری اولیه

در روش پیشنهادی برای برچسب‌گذاری اولیه پیکسل‌ها از چند تابع اکتشافی اولیه استفاده شده است که در ادامه توضیح داده می‌شود. برچسب‌های معنایی موجود به دو گروه اشیا و پس زمینه تقسیم می‌گردد. برچسب‌های معنایی موجود در گروه اشیا همانند انسان، ماشین و هواپیما که معنای شی ساختاریافته را دارند و برچسب‌های معنایی موجود در گروه پس‌زمینه همانند آب، آسمان، زمین و جاده است که حالت غیرساختاریافته را دارا می‌باشند. در ابتدا برای تصویر I_j، نواحی برجسته[1] آن از طریق الگوریتم [۱۸] بدست می‌آید. سپس برای هر ناحیه تصویر $\{r_k^j\}_{k=1}^{s_j}$، مقدار برجستگی از طریق میانگین گیری مقدار برجستگی پیکسل‌های موجود در ناحیه به دست می‌آید. نواحی از تصویر که دارای مقدار برجستگی بزرگتر از یک حد آستانه می‌باشند، برچسب‌های معنایی از t_j که ساختاریافته بوده و جز گروه اشیا هستند اختصاص می‌یابد و برای مابقی نواحی که مقدار برجستگی‌آن‌ها از یک حد آستانه پایینتر است، برچسب‌های گروه پس زمینه اختصاص می‌یابد. برچسب گذاری اولیه برای تمامی تصاویر انجام می‌پذیرد. پس از تعیین برچسب اولیه برای هر پیکسل تصویر، برای هر گروه معنایی یک مدل متمایزکننده آموزش می‌بینند. برای آموزش مدل مذکور در روش پیشنهادی از ماشین بردار پشتیبان استفاده شده است. همانطور که توضیح داده شده است، L بیانگر تعداد برچسب‌های معنایی موجود در مجموعه تصاویر است. برای هر برچسب معنایی همانند l، نواحی به هم پیوسته از هر تصویر که دارای برچسب l است استخراج می‌گردد. سپس نواحی استخراج شده با

[1] Salient Region

استفاده از کیفی از ویژگی‌های SIFT، رنگ و بافت توصیف می‌گردد. این رویه برای تمامی برچسب‌های معنایی انجام می‌گردد. سرانجام، برای آموزش مدل برای گروه معنایی l، توصیفگرهای به دست آمده برای برچسب معنایی l، به عنوان نمونه‌های آموزشی مثبت و نواحی مربوط به گروه‌های معنایی به غیر از l، به عنوان نمونه‌های آموزشی منفی در نظر گرفته می‌شوند. سپس با استفاده از رابطه‌ی ماشین بردار پشتیبان که در ادامه آورده می‌شود مدل آموزش می‌بینند [۱۹]:

$$\begin{aligned} \min_{m_l,b} \quad & \frac{1}{2} m_l^T m_l \\ s.t. \quad & +1\times(m_l^T d_{i_j(l)} + b) \geq 1 \\ & -1\times(m_l^T d_{i_j(\neg l)} + b) \geq 1 \\ & \forall\, j = 1,\ldots,N \\ & ,\, l \in \{1,\ldots,L\} \end{aligned} \qquad (۴)$$

که m_l بیانگر مدل آموزش دیده برای برچسب معنایی lام می‌باشد. همچنین برای آموزش مدل متن هر تصویر نیز از ایده مشابهی استفاده می‌گردد. مدل متن هر تصویر، چیدمان مکانی برچسب‌های معنایی آن در نظر گرفته می‌شود. بدین منظور در تصاویری که گروه معنایی l، گروه معنایی برجسته در این تصاویر می‌باشد، به عنوان نمونه‌های آموزشی مثبت در نظر گرفته شده و مابقی تصاویر به عنوان نمونه‌های آموزشی منفی در نظر گرفته می‌شود. توصیفگر متنی بیانگر توصیف چیدمان مکانی برچسب‌های معنایی هر تصویر می‌باشد. بدین منظور، در این مقاله از ایده انطباق هرمی مکانی SPM [۲۰] بهره شده است. در ایده مذکور تصویر به چندین سطح تقسیم می‌گردد. سپس برای هر سلول به دست آمده از تقسیم بندی، هیستوگرام تعداد رخداد برچسب‌های معنایی مورد محاسبه قرار می‌گیرد. در نهایت با استفاده از رویه مشابهی(رابطه‌ی ۴)، مدل توصیفگر متن برای هر گروه معنایی آموزش می‌بینند.

۲-۲ بهینه سازی

همانطور که پیشتر نیز ذکر شد، رابطه‌ی ۴، رابطه‌ی غیرمحدب است و در این مقاله برای بهینه سازی آن از یک روش تکراری مبتنی بر حرکت استفاده شده است. الگوریتم مبتنی بر حرکت پیشنهادی نیز مبتنی بر بسط است. ایده الگوریتم بهینه سازی در روش پیشنهای از الگوریتم بهینه سازی بسط-α [۲۱] بهره گرفته شده است. در ادامه رابطه‌ی ۴ به گونه‌ای بازنویسی می‌شود که بتوان با استفاده از روش پیشنهادی ارائه شده آن را بهینه نمود. برای بازنویسی آن، رابطه‌ی ۴ براساس هر برچسب معنایی α تشکیل خواهد شد. در ادامه، هر عبارت رابطه‌ی ۴ را به صورت جداگانه مورد بررسی و تحلیل قرار می‌دهیم. عبارت یک رابطه‌ی ۴ به صورت زیر بازنویسی می‌شود:

$$J_1(\alpha) = ((l' - \beta h l'')D - (l' - \beta h l'')M)^T \\ ((l' - \beta h l'')D - (l' - \beta h l'')M) \qquad (۵)$$

که l' بیانگر ماتریسی $1\times(S\times L)$ است که i امین درایه آن بیانگر این است که برچسب $\mathrm{mod}(i,S)$ امین ناحیه، در صورتی که مقدار i امین درایه یک باشد، $\lfloor i/S \rfloor + 1$ می‌باشد. متغیر β بیانگر ماتریسی $1\times S$ است که هر درایه این ماتریس می‌تواند مقدار صفر یا یک داشته باشد که مقدار یک نمایانگر این است که ناحیه مذکور در برچسب‌گذاری مقدار α را خواهد گرفت و اگر مقدار صفر داشته باشد، برچسب پیکسل همان مقدار قبلی باقی می‌ماند. ماتریس h نیز به صورت زیر تعریف می‌گردد:

$$h = \underset{i=1}{\overset{N}{join}}(rep(I_{s_i\times s_i}, L)) \qquad (۶)$$

که $I_{s_i\times s_i}$ بیانگر ماتریسی همانی با اندازه‌ی $s_i \times s_i$ است. تابع $rep(s,n)$ ماتریس s را به صورت ستونی n بار تکرار می‌نماید و همچنین تابع $join$ ، مجموعه ماتریس‌های ورودی را به صورت قطری به یکدیگر الحاق می‌نماید و ماتریس نهایی را به عنوان خروجی در اختیار قرار می‌دهد. ماتریس l'' در رابطه‌ی ۵ نیز به صورت زیر تعریف می‌گردد:

$$l'' = diag(l' - f(l,l')) \qquad (۷)$$

که تابع $diag(.)$ ، به عنوان ورودی بردار را گرفته و ماتریسی قطری را به عنوان خروجی برمی‌گرداند که در آن درایه های بردار ورودی بر روی قطر ماتریس خروجی قرار گرفته‌اند. تابع f به عنوان خروجی ماتریسی به ابعاد $1\times(L\times S)$ برمی‌گرداند که هر درایه این ماتریس به صورت زیر تعریف می‌گردد:

$$f_{1,(i,s)}(l,l') = \begin{cases} -1 & if\ \alpha == l \\ l'(1,(i-1)*S+s) & otherwise \end{cases} \qquad (۸)$$

به عبارت دیگر رابطه‌ی $(l' - \beta h l'')$ ، بیانگر برچسب‌های هر پیکسل پس از اعمال متغیر β است که در آن پیکسل‌هایی که مقدار β متناظر آن صفر است، برچسب پیکسل مقدار قبلی باقی می‌ماند و در صورتی که مقدار β متناظر آن یک باشد برچسب پیکسل α خواهد ماند. رابطه‌ی ۵ سعی می‌نماید که متغیر l را به گونه‌ای به دست آورد که نحوه ظهور نواحی از تصویر که به یک برچسب معنایی مشترک اختصاص یافته‌اند مشابه باشد. ماتریس D نیز به صورت زیر تعریف شده است:

$$D=\begin{bmatrix} \overbrace{\begin{matrix} D_1 & 0 & \cdots & \cdots & \cdots & 0 & 0 \\ 0 & D_2 & & & & & 0 \\ & \ddots & & & & & \\ \vdots & & D_N & & & & \vdots \\ \vdots & & & \ddots & & & \vdots \\ \vdots & & & & D_1 & & \vdots \\ & & & & & D_2 & \\ 0 & & & & & \ddots & 0 \\ 0 & 0 & \cdots & \cdots & \cdots & 0 & D_N \end{matrix}}^{N} \end{bmatrix}^{N\times L} \quad (۹)$$

که D_i به صورت زیر تعریف می‌شود:

$$D_i=\begin{bmatrix} d_i(1) \\ d_i(2) \\ \vdots \\ d_i(s_i) \end{bmatrix} \quad (۱۰)$$

که $d_i(j)$ بیانگر توصیفگر ناحیه jام از تصویر iام است. همچنین ماتریس M نیز به صورت زیر تعریف شده است:

$$M=\begin{bmatrix} \overbrace{\begin{matrix} m_1 & 0 & \cdots & \cdots & \cdots & 0 & 0 \\ 0 & m_1 & & & & & 0 \\ & \ddots & & & & & \\ \vdots & & m_1 & & & & \vdots \\ \vdots & & & \ddots & & & \vdots \\ \vdots & & & & m_L & & \vdots \\ & & & & & m_L & \\ 0 & & & & & \ddots & 0 \\ 0 & 0 & \cdots & \cdots & \cdots & 0 & m_L \end{matrix}}^{N} \end{bmatrix}^{N\times L} \quad (۱۱)$$

که m_i بیانگر مدل برچسب معنایی iام است. همینطور برای در نظر گرفتن اطلاعات معنایی در روش پیشنهادی، عبارت دوم رابطه‌ی ۳ نیز به صورت زیر بازنویسی می‌شود:

$$J_2(\alpha)=((l'-\beta hl'')C-(l'-\beta hl'')M_C)^T \quad (۱۲)$$
$$((l'-\beta hl'')C-(l'-\beta hl'')M_C)$$

که در ماتریس C، اطلاعات تعداد پیکسل‌های موجود در هر ناحیه به گونه ای ذخیره شده است که عبارت $(l'-\beta hl'')C$ بیانگر توصیفگر معنایی تصاویر با استفاده از برچسب‌های اختصاص داده شده به آن است. بایستی توجه گردد که ماتریس‌های C و M_C نیز مشابه ماتریس‌های D و M ساخته می‌شود. ماتریس M_C نیز بیانگر مدل توصیفگر متن هر گروه معنایی است که همانطور که در بخش ۳ توضیح داده شده است با استفاده از ماشین بردار پشتیبان و برچسب‌های معنایی به دست آمده برای پیکسل‌ها به دست آمده است. در نهایت تابع بهینه سازی به صورت زیر تبدیل می‌گردد:

$$\min_{\beta}(((l'-\beta hl'')D-(l'-\beta hl'')M)^T((l'-\beta hl'')D-(l'-\beta hl'')M)$$
$$+((l'-\beta hl'')C-(l'-\beta hl'')M_C)^T((l'-\beta hl'')C-(l'-\beta hl'')M_C)) \quad (۱۳)$$

با توجه به اینکه رابطه‌ی ۱۳، یک تابع محدب است، بهینه سازی آن به راحتی انجام می‌پذیرد. در این مقاله برای بهینه کردن رابطه‌ی ۱۳ از بسته‌ی بهینه سازی CVX [۲۲] استفاده شده است. در روش پیشنهادی در هر بار تکرار روش، به ازای هر برچسب معنایی رابطه‌ی ۱۳ تشکیل شده و بهینه سازی می‌گردد. هنگامی که رابطه‌ی ۱۳ به ازای تمامی برچسب‌های معنایی تشکیل گردد و بهینه شود، یک دور از الگوریتم بهینه سازی انجام پذیرفته است. در صورتی که در دو دور متوالی، برچسب‌های معنایی هر پیکسل تغییری نیابد، همگرایی رخ داده است و الگوریتم متوقف شده و خروجی نهایی گزارش می‌گردد. بایستی دقت گردد که روش بهینه سازی پیشنهای همانند الگوریتم بهینه سازی بسط-α عمل می‌نماید که در روش پیشنهادی توانسته است پاسخ‌های قابل قبولی را در اختیار قرار دهد.

بایستی دقت گردد که در فاز آموزش، پس از هر دور به روزرسانی برچسب پیکسل‌ها، مدل ظهور هر برچسب معنایی و همچنین مدل اطلاعات متنی آن، باری دیگر با استفاده از ماشین بردار پشتیبان آموزش می‌بینند.

۳ نتایج

در این بخش، نتایج روش پیشنهادی بر روی تصاویر مجموعه داده‌ی استاندارد MSRC ارائه شده است. این مجموعه داده شامل ۲۱ گروه معنایی می‌باشد. تصاویر پایگاه داده‌ی MSRC به صورت استاندارد به دو دسته‌ی آموزش و آزمایش تقسیم شده است که در این مقاله ما نیز از تقسیم‌بندی استاندارد استفاده کرده‌ایم. تعداد کل تصاویر آموزشی و آزمایشی به ترتیب ۲۷۵ و ۳۱۵ تصویر می‌باشد. دسته‌های موجود در این مجموعه داده شامل گاو، گوسفند، سبزه، درخت، گل، جاده، آسمان، علامت (تابلوهای راهنمایی، تابلوهای کمکی)، آب، پرنده، نیمکت، سگ، گربه، چهره انسان، بدن انسان، هواپیما، دوچرخه، خودرو، کشتی و کتاب می‌باشد. بسیاری از اشیا موجود در این مجموعه داده ساختار یافته می‌باشند.

برای بدست آوردن توصیفگر هر ناحیه در روش پیشنهادی، از کیفی از ویژگی‌ها [۲۳] استفاده شده است. بدین منظور، در ابتدا، نقاط کلیدی به صورت متراکم با قدم نمونه‌برداری ۸ پیکسل در جهت افقی و عمودی نمونه برداری شده است. برای توصیف هر ناحیه از SIFT [۲۴]، توصیفگر رنگ Hue [۲۵] و بانک فیلتر استفاده شده است. برای هر توصیفگر به طور مستقل فرهنگ لغت

با ۲۰۰ کلمه ساخته شده است. سپس، هیستوگرام‌های هر توصیفگر با یکدیگر الحاق شده‌اند تا توصیفگر نهایی هر ناحیه به دست آید.

در این مقاله و در کارهای مشابه در حوزه‌ی بخش‌بندی معنایی برای ارزیابی کار، از میانگین دقت به ازای گروه (PC[1]) استفاده می‌گردد [۲۶]. در معیار PC، ابتدا نسبت پیکسل‌های درست شناسایی شده برای هر گروه معنایی محاسبه می‌گردد و سپس میانگین گرفته می‌شود. بدین منظور ابتدا برای برای هر گروه معنایی، نسبت پیکسل‌های درست شناسایی شده، با استفاده از رابطه‌ی زیر محاسبه می‌گردد:

$$C_{ij} = \sum_{z \in I} \delta(I_G(z) = 1)\delta(I_p(z) = j) \qquad (۱۴)$$

که I بیانگر تصویر اصلی، I_G بیانگر تصویر برچسب زده‌شده‌ی صحیح و I_p بیانگر تصویر برچسب زده‌شده‌ی پیش‌بینی شده می‌باشد. همچنین z بیانگر پیکسل موجود در تصویر است. سپس میانگین دقت به ازای گروه به صورت زیر محاسبه می‌گردد:

$$PC = \frac{1}{L}\sum_{i}\frac{C_{ii}}{\sum_{j} C_{ij}} \qquad (۱۵)$$

در این مقاله، جهت ارائه نتایج و ارزیابی مناسب روش پیشنهادی از ارزیابی‌های کمی و بصری استفاده شده است. به همین دلیل در ابتدا نتایج روش به صورت کیفی و بصری بر روی چندین تصویر، ارائه و تحلیل می‌گردد. سپس در گام بعدی بر روی هر مجموعه گروه معنایی موجود در مجموعه داده‌ی استاندارد MSRC، دقت بخش‌بندی معنایی هر گروه و میانگین دقت به ازای گروه گزارش می‌گردد.

۳-۱ نتایج بصری

در این بخش نتایج بصری روش پیشنهادی بر روی مجموعه‌ای از تصاویری که دارای برچسب معنایی مشابهی هستند ارائه می‌گردد. بدین منظور در شکل ۲-الف، سه تصویر که هر سه دارای گروه معنایی پرنده و سبزه می‌باشند به سیستم ارائه گردیده است و نتایج حاصل از آن در سطر دوم شکل ۲-الف نشان داده شده است. همانطور که در شکل نیز دیده می‌شود، روش پیشنهادی به خوبی توانسته است، نواحی مربوط به گروه‌های معنایی یکسان را از یکدیگر تمیز داده و به یک شی یکسان نسبت دهد. روش پیشنهادی در شکل ۲-ب بر روی مجموعه تصاویر دیگری که دارای مجموعه برچسب‌های سگ، سبزه و جاده هستند نمایش داده شده است. همانگونه که در شکل نیز مشاهده می‌گردد، روش پیشنهادی به خوبی تونسته است نتایج قابل قبولی را تولید نماید. اگر چه در تصویر سمت راست شکل ۲-ب، بخشی از پس‌زمینه نیز با شی سگ تلفیق شده است. علت این امر می‌تواند به خاطر بخش‌بندی اولیه تصویر باشد. در روش پیشنهادی همانگونه که ذکر شد، در ابتدا تصویر با استفاده از الگوریتم بخش‌بندی [۷] به چندین قسمت تقسیم می‌گردد. در صورتی که بخش‌بندی اولیه دارای خطا باشد، خطای حاصل از آن در طی کلیه قدم‌های بعدی منتشر می‌گردد. همچنین در شکل ۲-ج بر روی مجموعه تصاویری با برچسب علامت اعمال شده است. در شکل ۲-د نیز بر روی مجموعه تصاویر با برچسب گوسفند اعمال شده است و روش توانسته است که دقت قابل قبولی را به دست آورد و نواحی متناظر را به دست آورد.

در شکل ۳، نتایج روش پیشنهادی با روش‌های [۲۷] و [۹] بر روی چندین تصویر آزمون مجموعه داده‌ی MSRC مورد مقایسه قرار گرفته است. نتایج روش [۲۷]، از مقاله آن‌ها گزارش شده است و نتایج روش [۹] نیز از اجرای برنامه آن‌ها که در وبسایت در دسترس عموم بوده است، به دست آمده است. همانگونه که در شکل ۳-الف نشان داده شده است، روش پیشنهادی توانسته اطراف شی را به صورت هموارتری تشخیص دهد. در شکل ۳-ب، روش پیشنهادی بخش بزرگی از تصویر راکه شامل سبزه است را به درستی تشخیص داده است. درحالیکه روش [۲۷]، ناحیه مذکور را جاده تشخیص داده است و روش [۹] شی موجود در تصویر را به اشتباه تشخیص داده است. همچنین بایستی دقت گردد که خروجی روش پیشنهادی به بخش‌بندی اولیه‌ی تصویر نیز وابسته می‌باشد. در تصویر ۳-د، در بخش‌بندی اولیه، قسمتی از ساختمان و درخت در یک بخش قرار گرفته‌اند، در نتیجه، به هنگام برچسب زنی، کل بخش یک برچسب معنایی خواهد داشت. این امر موجب می‌گردد که خطای موجود در بخش‌بندی اولیه، در نتایج روش پیشنهادی نیز تاثیرگذار باشد.

۳-۲ نتایج کمی

نتایج کمی روش پیشنهادی بر روی مجموعه داده‌ی MSRC در جدول ۱ نشان داده شده است. همانطور که در جدول نشان داده شده است، در برخی از گروه‌های معنایی همانند گل، صورت انسان، خودرو، گوسفند جواب قابل قبولی را ایجاد کرده است. اما در برخی از گروه‌ها همانند دوچرخه و یا انسان نتوانسته است خیلی خوب عمل نماید. با تحلیل عملکرد روش پیشنهادی، علت اصلی کارکرد نسبتا ضعیف روش در گروه‌های معنایی مذکور در توصیفگرهای نواحی است. با توجه به اینکه هر ناحیه در روش پیشنهادی با استفاده از هیستوگرامی از کلمات بصری توصیف می‌گردد. در برخی از مواقع به دلیل اینکه برخی از نواحی به درستی توصیف نمی‌گردند، باعث ایجاد اشتباهاتی می‌شود. در گروه‌های معنایی دوچرخه و یا انسان، بخش‌بندی اولیه تصاویر باعث می‌گردد که نواحی کوچکی در تصویر ایجاد گردد که در اغلب موارد نواحی مذکور دارای توصیفگرهای دقیقی نبوده و به همین دلیل باعث کاهش دقت می‌گردد. بایستی توجه گردد که در روش پیشنهادی، مجموعه تصاویر آزمایشی ابتدا به ۲۱ دسته خوشه‌بندی

[1] Per Class (PC)

جدول ۱ نرخ شناسایی پیکسل برای هر دسته از شی در مجموعه داده MSRC-21

گروه معنایی	گاو	سبزه	درخت	گل	ساختمان	علامت	آسمان	پرنده	جاده	آب	کتاب	صندلی	کشتی	سگ	گربه	انسان	صورت انسان	خودرو	دوچرخه	هواپیما	گوسفند	میانگین
[۹] (۲۰۱۱)	۸۱	۸۳	۷۰	۸۲	۱۲	۶۹	۸۴	۵۱	۶۶	۵۵	۶۱	۵۹	۵۸	۴۴	۵۳	۹	۹۷	۸۷	۹۲	۹۱	۹۳	۶۷
[۲۸] (۲۰۱۲)	۱۰	۹۲	۴۹	۹۸	۷۰	۸۸	۸۳	۴۸	۷۵	۸۲	۹۸	۷۰	۲۳	۷۶	۹۵	۴۳	۶۲	۲۰	۵۲	۳۶	۱۰	۶۱
[۲۹] (۲۰۱۴)	۶۱	۹۳	۸۴	۸۱	۷۴	۷۱	۸۰	۳۶	۷۵	۷۵	۷۲	۲۵	۱۰	۳۹	۵۲	۴۹	۷۵	۶۲	۷۵	۵۵	۶۰	۶۲
[30] (۲۰۱۵)	۹۶	۵۴	۵۵	۱۰۰	۲۰	۹۹	۶۱	۹۵	۲۶	۴۰	۱۰۰	۹۹	۶۲	۹۷	۱۰۰	۱۰	۷۳	۷۳	۹۷	۵۷	۸۵	۷۱
[۳۱] (۲۰۱۵)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	۷۳
[۳۲] (۲۰۱۶)	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	۶۹
روش پیشنهادی	۶۵	۹۲	۷۴	۱۰۰	۶۱	۷۴	۷۸	۷۲	۷۶	۴۵	۱۰۰	۹۳	۹۶	۸۳	۶۱	۲۴	۸۹	۶۱	۳۹	۶۵	۶۴	۷۲

می‌گردند، سپس تصاویر هر خوشه به روش پیشنهادی داده می‌شود و نتایج به دست می‌آید. نتایج روش پیشنهادی با روش‌های [۹، ۲۸-۳۲] مورد بررسی قرار گرفته است. روش پیشنهادی به طور میانگین توانسته است، ۵ درصد افزایش دقت نسبت به روش [۹] داشته باشد. همچنین روش پیشنهادی در بسیاری از گروه‌های معنایی توانسته است دقت قابل قبولی را در مقایسه با دیگر روش‌ها به دست آورد. در برخی از گروه‌ها نیز همانند دوچرخه، روش‌های دیگر بهتر از روش پیشنهادی عمل کرده است. دقت روش [۳۱] به طور میانگین ۱% بیشتر از روش پیشنهادی می‌باشد. بایستی توجه نمود که در روش [۳۱]، از یادگیری انتقالی هدایتی استفاده شده است. در نتیجه، تمامی تصاویر آزمایشی نیز در طی مرحله آموزش مورد استفاده قرار گرفته‌اند. در حالی که در مابقی روش‌ها، تصاویر آزمایشی به هیچ عنوان در طی مرحله آموزش مورد استفاده قرار نگرفته‌اند.

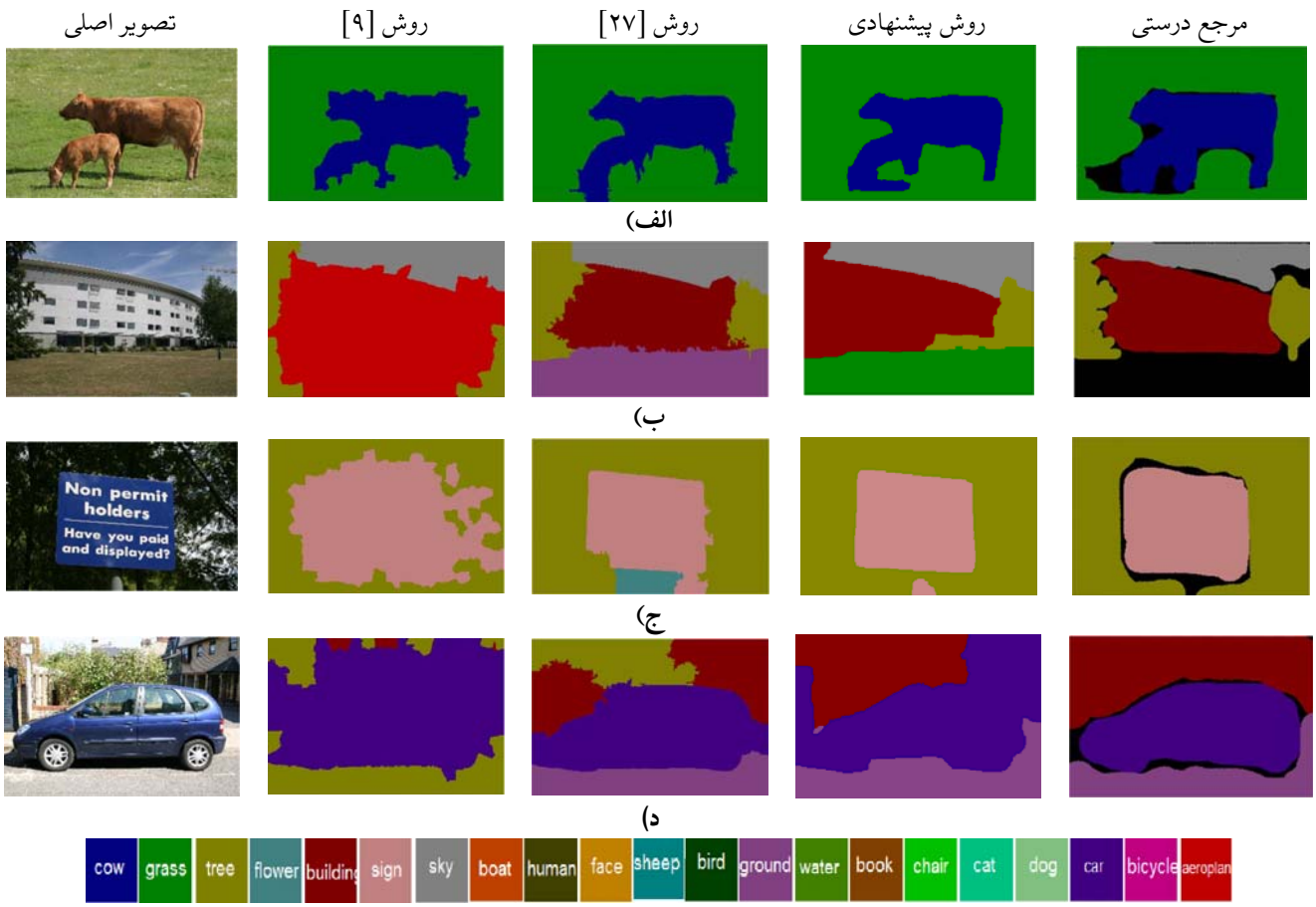
برای ارزیابی پیچیدگی محاسباتی روش پیشنهادی، پیچیدگی زمانی بهینه سازی تابع هزینه پیشنهادی در نظر گرفته می‌شود. در این قدم، زمان اجرایی روش پیشنهادی برای ۲۰۰ تصویر متفاوت محاسبه می‌شود. نتایج بیانگر این است که روش پیشنهادی به طور میانگین به ۱۹ ثانیه (با انحراف معیار تقریبی ۲ ثانیه) زمان نیاز دارد تا تصویر را برچسب گذاری نماید. روش پیشنهادی با استفاده از MATLAB پیاده سازی شده است و بر روی واحد مرکزی محاسبات یک هسته ای ۲/۶ گیگاهرتز با ۸ گیگابایت حافظه اجرا شده است.

۴ نتیجه گیری

در این مقاله، یک روش پیشنهادی برای بخش‌بندی معنایی تصاویری با برچسب‌های متنی، ارائه شد. در روش پیشنهادی، با استفاده از روش ارائه شده در این مقاله، برچسب‌های معنایی تصاویر به دست آمد. سپس با بهینه‌سازی تابع هزینه پیشنهادی، برچسب‌های معنایی هر پیکسل به دست آمد. یکی از مهمترین ویژگی روش پیشنهادی ارائه شده این امر است که روش پیشنهادی در بخش بندی معنایی با داده‌های نظارتی ضعیف توانسته است، اطلاعات سطح شی و سطح متن را به صورت صریح و همزمان در طی مرحله استنتاج دخیل نماید. روش‌های پیشین ارائه شده در این حوزه، برای استفاده از اطلاعات سطح بالا، تنها از اطلاعات نواحی مشابه در تصاویر متفاوت بهره گرفته‌اند که نواحی مذکور اغلب تنها قسمتی از یک شی را نمایش می‌دادند. روش پیشنهادی، در حضور برخی اشیا، کارایی پایینی دارد که علت آن نیز توصیف غیردقیق ناحیه‌های کوچک در روش پیشنهادی است. روش پیشنهادی بر روی مجموعه MSRC اعمال گردیده

است و نتایج حاصل گزارش شده است. نتایج حاصل بیانگر کارکرد مناسب روش پیشنهادی است.

شکل۲ نمونه‌هایی از نتایج کیفی روش پیشنهادی بر روی مجموعه داده‌ی MSRC.



شکل۳ مقایسه نتایج کیفی روش پیشنهادی با برخی از روش‌های موجود بر روی مجموعه داده‌ی MSRC.

مراجع

[1] C. Liu, J. Yuen, and A. Torralba, "Nonparametric Scene Parsing via Label Transfer," *Pattern Analysis and Machine Intelligence,* vol. 33, pp. 2368-2382, 2011.

[2] C. M. Bishop, *Pattern recognition and machine learning*. New York: Springer, 2006.

[3] S. Gould, "Probabilistic Models for Region-Based Scene Understanding," Doctor of Philosophy, Electrical Engineering Department Stanford University, 2010.

[4] M. P. Kumar and D. Koller, "Efficiently Selecting Regions for Scene Understanding," in *Computer Vision and Pattern Analysis*, pp. 3217-3224, 2010.

[5] L. Ladicky, "Global Structured Models towards Scene Understanding," Doctor of Philosophy, Oxford Brookes University, 2011.

[6] J. Yao, S. Fidler, and R. Urtasun, "Describing the Scene as a Whole: Joint Object Detection, Scene Classification and Semantic Segmentation," in *Computer Vision and Pattern Recognition*, pp. 702-709, 2012.

[7] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik, "Contour Detection and Hierarchical Image Segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 33, pp. 898-916, 2011.

[8] J. J. Corso, "Toward Parts-Based Scene Understanding with Pixel-Support Parts-Sparse Pictorial Structures," *Pattern Recognition Letters,* vol. 34, pp. 762-769, 2013.

[9] A. Vezhnevets, V. Ferrari, and J. M. Buhmann, "Weakly Supervised Semantic Segmentation with a Multi-Image Model," in *IEEE International Conference on Computer Vision (ICCV)*, pp. 643-650, 2011.

[10] B. Alexe, T. Deselaers, and V. Ferrari, "What is an object?," in *IEEE Computer Vision and Pattern Recognition (CVPR)*, pp. 73-80, 2010.

[11] Y. Liu, J. Liu, Z. Li, J. Tang, and H. Lu, "Weakly-Supervised Dual Clustering for Image Semantic Segmentation," in *IEEE Computer Vision and Pattern Recognition (CVPR)*, pp. 2075-2082, 2013.

[12] G. Kim, E. P. Xing, L. Fei-Fei, and T. Kanade, "Distributed cosegmentation via submodular optimization on anisotropic diffusion," in *International Conferance on Computer Vision (ICCV)*, pp. 169-176, 2011.

[13] A. Joulin, F. Bach, and J. Ponce, "Multi-class cosegmentation," in *IEEE Computer Vision and Pattern Recognition* (*CVPR*), pp. 542-549, 2012.

[14] F. Z. Xing, E. Cambria, W. B. Huang, and Y. Xu, " Weakly supervised semantic segmentation with superpixel embedding," in *IEEE International Conference on Image Processing* (*ICIP*), pp. 1269-1273, 2016

[15] J. Xu, A. G. Schwing, and R. Urtasun, "Learning to segment under various forms of weak supervision," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3781-3790, 2015.

[16] Y. Niu, Z. Lu, S. Huang, P. Han, and J. R. Wen, "Weakly Supervised Matrix Factorization for Noisily Tagged Image Parsing," in *International Joint Conference on Artificial Intelligence* (*IJCAI*), pp. 3749-3755, 2015.

[17] F. Saleh, M. S. A. Akbarian, M. Salzmann, L. Petersson, S. Gould, and J. M. Alvarez, "Built-in foreground/background prior for weakly-supervised semantic segmentation," in *European Conference on Computer Vision*, pp. 413-432, 2016.

[18] C. Yang, L. Zhang, H. Lu, M.-H. Yang, and X. Ruan, "Saliency Detection via Graph-Based Manifold Ranking," in *IEEE conference on Computer Vision and Pattern Recognition*, pp. 3166-3173, 2013.

[19] V. N. Vapnik, *The Nature of Statistical Learning Theory*: Springer-Verlag, 1995.

[20] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Computer Vision and Pattern Recognition*, pp. 2169–2678, 2006.

[21] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *IEEE Transactions on pattern analysis and machine intelligence,* vol. 23, pp. 1222-1239, 2001.

[22] M. Grant and S. Boyd, "cvx Users' Guide," 2012.

[23] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories," in *Computer Vision and Pattern Analysis*, pp. 2169-2178, 2006.

[24] D. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vision,* vol. 60, pp. 91–110, 2004.

[25] J. v. d. Weijer and C. Schmid, "Coloring Local Feature Extraction," in *European Conference on Computer Vision*, 2006.

[26] G. Csurka, D. Larlus, F. Perronnin, and F. Meylan, "What is a good evaluation measure for semantic segmentation?," in *British Machine Vision Conference* (*BMVC*), 2013.

[27] Y. Li, Y. Guo, Y. Kao, and R. He, "Image Piece Learning for Weakly Supervised Semantic Segmentation," *IEEE Transactions on Systems, Man, and Cybernetics: Systems,* 2016.

[28] S. Liu, S. Yan, T. Zhang, C. Xu, J. Liu, and H. Lu, "Weakly supervised graph propagation towards collective image parsing," *IEEE Trans. Multimedia,* vol. 14, pp. 361–373, 2012.

[29] E. Akbas and N. Ahuja, "Low-level hierarchical multiscale segmentation statistics of natural images," *IEEE Transaction on Pattern Analysis and Machine Intelligence* (*PAMI*), vol. 36, pp. 1900-1906, 2014.

[30] Y. Niu, Z. Lu, S. Huang, P. Han, and J. R. Wen, "Weakly supervised matrix factorization for noisily tagged image parsing," in *Proceedings of the 24th International Conference on Artificial Intelligence. AAAI Press*, pp. 3749-3755, 2015.

[31] J. Xu, A. G. Schwing, and R. Urtasun, "Learning to segment under various forms of weak supervision," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3781-3790, 2015.

[32] F. Z. Xing, E. Cambria, W. B. Huang, and Y. Xu, "Weakly supervised semantic segmentation with superpixel embedding," in *IEEE International Conference on Image Processing* (*ICIP*), pp. 1269-1273, 2016.

پروین رزاقی مدرک کارشناسی علوم کامپیوتر را از دانشگاه تبریز در سال ۸۶ اخذ نموده است. ایشان کارشناسی ارشد و دکتری تخصصی خود را در رشته‌ی مهندسی کامپیوتر گرایش هوش مصنوعی از دانشگاه صنعتی اصفهان به ترتیب در سال‌های ۸۹ و ۹۳ به اتمام رسانده است. هم اکنون ایشان به عنوان استادیار در دانشگاه تحصیلات تکمیلی علوم پایه زنجان، دانشکده علوم رایانه و فناوری اطلاعات مشغول به کار است. از علاقه‌مندی‌های ایشان می‌توان به یادگیری‌ماشین، شناسایی‌الگو، بینایی‌کامپیوتر و پردازش‌تصویر اشاره نمود.