بررسی روش های تشخیص اشیاء مبتنی بر منطقه

تاریخ دریافت:
۳ بهمن ماه ۱۴۰۲

تاریخ پذیرش:
٤ مهرماه ۱۴۰۲

پژمان غلام نژاد*[۱]، امیرمهدی سازدار[۱]، امیرحسین زنگنه[۱]
۱. دانشکده مهندسی رایانه و فناوری اطلاعات، دانشگاه علوم و فنون هوایی شهید ستاری

چکیده
تشخیص اشیاء، یک شیوه در پردازش تصویر و بینایی ماشین است که برای مکان‌یابی نمونه‌هایی از اشیاء، در تصاویر یا فیلم‌ها مورد استفاده قرار می‌گیرد. الگوریتم‌های تشخیص شیء، معمولاً از روش‌های یادگیری ماشین یا روش‌های یادگیری عمیق استفاده می‌نمایند. با توجه به این که شبکه‌های عصبی پیچشی مبتنی بر منطقه، خانواده‌ای از مدل‌های شبکه‌های عصبی پیچشی هستند که برای تشخیص اشیاء مورد استفاده قرار می‌گیرند، و دارای این خصوصیت می‌باشند که در حالتِ کاستی‌های طبقه‌بندی، علاوه بر استخراج ویژگی، اجزای بیشتری را نیز از قبل آموزش می‌دهند، در این مقاله به بررسی روش‌های تشخیص اشیاء مبتنی بر منطقه، با توجه به دسته‌بندی‌های اشاره شده، پرداخته می‌شود.

واژه‌های کلیدی: تشخیص اشیاء مبتنی بر پیش آموزشِ خود-سرپرست، تشخیص اشیاء مبتنی بر منطقه، شبکه¬های عصبی پیچشی مبتنی بر منطقه. تشخیص اشیاء مبتنی بر تعداد کمی عکس.

A survey of Region-based Object Detection models

Abstract
Object detection is a method in image processing and machine vision that is used to locate samples of objects in images or videos. Object detection algorithms usually use machine learning methods or deep learning methods. Considering that region-based convolutional neural networks are a family of convolutional neural network models that are used to detect objects, and have the characteristic that in the case of classification deficiencies, in addition to extracting feature, more components are also taught in advance, in this article, the method of detecting objects based on the region, according to the mentioned categories, is discussed. Also, the performance criteria used in the detection methods of mentioned objects are introduced.

Key words: Self-supervised pre-training object detection, Region-based object detection, Region-based Convolutional Neural Networks (R-CNN), Low-shot object detection.

ایمیل نویسنده مسئول: Pezhman.Gholamnezhad@gmail.com

۱-مقدمه

انسان به راحتی می‌تواند به اطراف خود نگاه نماید و هر چه را می‌بیند، به سرعت و با دقت تشخیص دهد. اما آموزش آن به کامپیوتر تا پایان دهه گذشته یک کار دشوار بوده است. این امر مستلزم شناسایی تمام نمونه‌های یک شیء (مانند ماشین‌ها، انسان‌ها، علائمِ خیابان‌ها و ...) در میدان دید است. به طور مشابه، سایر وظایف مانند طبقه‌بندی، تقسیم‌بندی، تخمینِ حرکت، درکِ صحنه و ...، مشکلات اساسی در بینایی کامپیوتر بوده اند. امروزه با پیشرفت دانش در حوزه هوش مصنوعی، بینایی ماشین به سیستم‌ها این امکان را می‌دهد که همانند انسان دنیای پیرامون خود را ببینند و درک کنند. تشخیص اشیاء، یک شیوه در پردازش تصویر و بینایی ماشین است که برای مکان‌یابی نمونه‌هایی از اشیاء، در تصاویر یا فیلم‌ها مورد استفاده قرار می‌گیرد. تشخیص و شناسایی اشیاء می‌تواند در کاربردهایی از قبیل تشخیص چهره[۱]، شمارش افراد، شناسایی اشیاء در صنایع، ماشین‌های خودران[۲]، ردیابی اشیاء[۳]، تشخیص هویت[۴]، تشخیص رفتار[۵]، شناسایی فعالیت‌های ناهنجار[۶]، و رباتیک مورد استفاده قرار گیرد.

مدل‌های تشخیص اشیاء اولیه، بر اساس مجموعه‌ای از استخراج‌کننده‌های ویژگی دست‌ساز ارائه شدند، از قبیل چارچوب تشخیص اشیاء ویُلا-جونز[۷]، که یک چارچوب تشخیص اشیاء مبتنی بر یادگیری ماشین با ارائه نرخ تشخیص بلادرنگ است[۱] و هیستوگرام گرادیان‌های جهت‌دار[۸][۲] که جهت‌گیری گرادیان را در بخش‌های محلی یک تصویر شمارش می‌کند. این مدل‌ها کند، نادرست بودند و در مجموعه داده‌های ناآشنا عملکرد ضعیفی داشتند. معرفی مجدد شبکه عصبی پیچشی و یادگیری عمیق برای طبقه‌بندی تصویر، چشم انداز ادراک بصری را تغییر داد[۳]. استفاده از آن در چالش تشخیص تصویر در مقیاس بزرگ مبتنی بر پایگاه داده تصویری ایمیج نِت[۹][۴]، الهام بخش تحقیقات بیشتر در مورد کاربرد آن در بینایی کامپیوتر شد.

اولین روش تشخیص اشیاء مبتنی بر یادگیری عمیق، استفاده از شبکه‌های عصبی پیچشی است[۵]. در این روش یک تصویر به عنوان ورودی دریافت شده و سپس به بخش‌های مختلف تقسیم می‌شود و هر بخش تصویر به عنوان یک تصویر جداگانه در نظر گرفته می‌شود. پس از تقسیم هر کدام از این بخش‌ها به کلاسِ مربوطه، تمام این بخش‌ها برای گرفتن تصویر اصلی با اشیاء ترکیب می‌شوند و مشخص می‌شود هر بخش دارای چه اشیایی است. این رویکرد کارآمد نبود، زیرا اشیاء موجود در تصویر می توانند نسبت‌های مختلف و در مکان‌های مختلف تصویر باشند. به عنوان مثال در برخی از موارد ممکن است بخش اعظم از یک تصویر باشند. گاهی اوقات هم ممکن است بخش کوچکی از تصویر باشند و اشکال مختلفی داشته باشند. در نتیجه به بخش‌های زیادی برای دسته‌بندی جداگانه نیاز است که این کار محاسبات زیادی را نیاز دارد.

پس از آن معماری شبکه پیچشی مبتنی بر منطقه توسط راس گیرشیک[۱۰] پیشنهاد شد[۶]. این معماری در سه مرحله عمل می‌کند: اول اشیاء را در کل تصویر را با استفاده از روش پیشنهاد منطقه[۱۱] استخراج نموده و سپس، ویژگی‌هایی را از هر یک مناطق استخراج می‌نماید. در نهایت تصویر را با استفاده از ماشین بردار پشتیبان، دسته-بندی می‌کند. در این روش با وجود این که نتایج بسیار دقیق بودند، اما آموزش شبکه با استفاده از مجموعه داده‌های مختلف موجود یک چالش بود. با این حال این رویکرد بعدها توسط راس گیرشیک، بهبود یافت و شبکه پیچشی مبتنی بر منطقه سریع[۱۲] نامیده شد[۷]. این رویکرد از یک الگوریتم پیشنهاد منطقه‌ای استفاده می‌نماید که جستجوی انتخابی را برای استخراج اشیاء مورد علاقه در صحنه مورد استفاده قرار می‌دهد. بدین ترتیب که از تمام تصویر به جای انتخاب یک ناحیه استفاده می-نماید. در اصل چهار بخش برای یک شی وجود دارد: مقیاس‌ها و اندازه-های مختلف، رنگ، بافت و محفظه. جستجوی انتخابی این الگوها را در تصویر مشخص می‌نماید و بر اساس آن مناطقی را پیشنهاد می‌دهد. در واقع در این روش، در ابتدا یک تصویر به عنوان ورودی دریافت می‌شود و بخش‌های جالب آن به کمک روش پیشنهادی مشخص می-شود. این بخش‌ها تغییر اندازه داده می‌شود تا متناسب با ورودی شبکه عصبی پیچشی باشد. شبکه عصبی پیچشی ویژگی‌ها را از هر بخش تصویر در می‌آورد و از ماشین بردار پشتیبان برای تقسیم‌بندی این بخش‌ها به دسته‌های مختلف استفاده می‌شود. در پایان از یک رگرسیون، برای پیش‌بینی محدوده‌ی هر شیء استفاده می‌شود. تمام این فرآیند باعث می‌شود که شبکه عصبی پیچشی مبتنی بر منطقه، خیلی کند شود. برای هر تصویر جدید این فرآیند ۴۰-۵۰ ثانیه طول می‌کشد که برای مجموعه داده‌های بزرگ این امر فرآیند رو دشوار و عملا غیر ممکن می‌نماید (ممکن است سال‌ها طول بکشد).

در سال ۲۰۱۶، معماری یُولو[۱۳] (شما فقط یک بار یک‌پارچه نگاه می‌کنید) توسط جوزف ردمون[۱۴] ارائه شد[۸]. در این معماری، تشخیص شیء، به عنوان یک مساله رگرسیون در کادرهای مرزی جدا شده از نظر فضایی و احتمالات کلاس مرتبط قاب می‌شوند. یک شبکه عصبی واحد، جعبه‌های محدود و احتمالات کلاس را مستقیماً از تصاویر کامل در یک ارزیابی پیش‌بینی می‌کند.

[1] Face recognition
[2] Slef-driving vehicles
[3] Object tracking
[4] Identification
[5] Behavior detection
[6] Abnormal activity detection
[7] Viola-Jones framework
[8] Histogram of Oriented Gradients (HOG)
[9] Imagenet
[10] Ross Girshick
[11] Region proposal method
[12] Fast RCNN
[13] YOLO (You Only Look Once)
[14] Joseph Redmon

با پیروی از معماری یُولو، روش شبکه عصبی پیچشی مبتنی بر منطقه سریع‌تر[1]، ارائه شد[۹]. این روش می‌تواند به طور موثر با استفاده از داده‌های آموزشی اضافه شده به ویژگی شبکه پیشنهادی منطقه[2] آموزش داده شود. شبکه پیشنهادی منطقه، اشیاء را بر اساس امتیاز شیئی نسبی آن‌ها خروجی می‌دهد. اشیاء استخراج شده از شبکه‌های پیشنهادی منطقه، توسط نظرسنجی رُلی[3] به اندازه یکسان در می‌آیند و تصاویر هم اندازه، توسط شبکه عصبی کاملاً متصل دسته‌بندی می‌شوند. این الگوریتم برای استخراج تمام اشیاء نیاز دارد که یک تصویر را چندین بار پیمایش نماید. با توجه به توالی سیستم‌ها عملکرد هر مرحله وابستگی زیادی به مرحله قبلی دارد.

روش‌های تشخیص اشیاء سنتی مبتنی بر مجموعه داده‌های تشخیص اشیاء نظارت شده بزرگی مانند پاسکال وُک[4] و اِم اِس کُوکُو[5] هستند که بیش از هزاران مثال مشروح در هر دسته شی دارند. با این حال، برچسب زدن داده ها اغلب گران و زمان‌بر است. این امر به‌ویژه در مورد تشخیص شی و تقسیم‌بندی نمونه، که نیازمند برچسب‌گذاری متراکم جعبه‌ها/ماسک‌های محدودکننده برای هر شی است، صادق است، فرآیندی که کندتر است و به آموزش حاشیه‌نویس بیشتری نسبت به طبقه‌بندی اشیاء نیاز دارد. علاوه بر این، برای کاربردهای تشخیص اشیای ریز دانه مانند شناسایی گونه‌های گیاهی یا جانوری، مجموعه داده‌های از پیش برچسب‌گذاری‌شده ممکن است وجود نداشته باشند، و ممکن است برچسب‌ها در محل توسط حاشیه‌نویسان متخصص جمع‌آوری شوند. برای حل این مشکلات، روش‌های تشخیص شیء بر مبنای تعداد کمی عکس تلاش می‌کنند تا کلاس‌های شی جدید (دیده‌نشده) را تنها بر اساس چند مثال، پس از آموزش بر روی بسیاری از نمونه‌های برچسب‌گذاری شده از کلاس‌های پایه (دیده‌شده) تشخیص دهند[۱۰]. تا همین اواخر، رویکرد استاندارد در تشخیص اشیاء مبتنی بر تعداد کمی عکس، یک پیش‌آموزش برای طبقه‌بندی، و سپس آموزشِ آشکارساز شی در کلاس‌های پایه، و در نهایت تنظیم دقیق در کلاس‌های جدید بود. با پیشرفت روش‌های یادگیری بازنمایی‌های خود نظارت، روش‌های تشخیص شیء مبتنی بر تعداد کمی عکس، از پیش‌آموز خود نظارت استفاده می‌کنند. دفاع از آشکارسازهای شی در برابر پ□چ[6] دشمن، و تشخیص شی نظامی در دفاع با استفاده از شبکه‌های کپسول چند سطحی، نمونه‌هایی از کاربردهای اخیر این موضوع در دفاع سامانه‌های الکترونیکی می‌باشند.

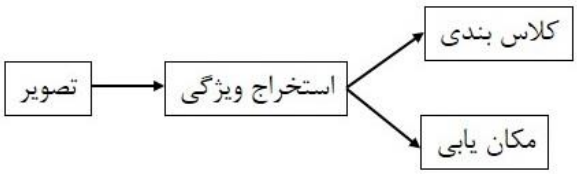
در این مقاله به بررسی معماری‌های مختلف تشخیص اشیاء، بر مبنای روش‌های مبتنی بر منطقه، با توجه به دسته‌بندی‌های اشاره شده، پرداخته می‌شود. در بخش ۲، مفاهیم پایه بیان می‌شود.

۲-مفاهیم پایه

در این بخش مفاهیم پایه در تشخیص اشیاء مرور می‌شود.

۲-۱-تشخیص اشیاء یک-مرحله‌ای[7]

مدل‌های تشخیص شی یک-مرحله‌ای، به دسته‌ای از مدل‌های تشخیص اشیا اشاره می‌کنند که از مرحله پیشنهاد منطقه مدل‌های دو مرحله‌ای عبور می‌کنند و تشخیص را مستقیماً بر روی نمونه‌گیری متراکم از مکان‌های از پیش تعریف‌شده از روی نقشه ویژگی[8]، با امکان اصلاح بعدی مکان‌های جعبه و نسبت‌های تصویر، پیش‌بینی می‌کنند. این نوع مدل‌ها معمولاً استنتاج سریع‌تری دارند، مانند: روش یُولُو[۱۱]، آشکارساز تک-تصویر[9][۱۲]، شبکه رِتینا[10][۱۳]، شبکه مرکز[11][۱۴]. شکل ۱ این ساختار را نمایش می‌دهد:



شکل ۱ - ساختار تشخیص اشیاء تک-مرحله‌ای

۲-۲-تشخیص اشیاء دو-مرحله‌ای[12]

در تشخیص شیء دو-مرحله‌ای، یک مدل برای استخراج نواحی از اشیاء و مدل دوم برای طبقه‌بندی و اصلاح بیشتر محلی‌سازی شیء، مورد استفاده قرار می‌گیرد. چنین روش‌هایی نسبتاً کند، اما بسیار قدرتمند شناخته شده‌اند، اما پیشرفت‌های اخیر مانند اشتراک‌گذاری ویژگی‌ها، آشکارسازهای ۲ مرحله‌ای را بهبود بخشید تا هزینه محاسباتی مشابهی با آشکارسازهای تک مرحله‌ای داشته باشند. شکل ۲ این ساختار را نمایش می‌دهد:

شکل ۲ - ساختار تشخیص اشیاء دو-مرحله‌ای

۲-۳-معماری‌های ستون فقرات[13]

شبکه ستون فقرات یک استخراج کننده ویژگی است که یک تصویر آر جی بی[14] را به عنوان ورودی می‌گیرد و یک یا چند نقشه ویژگی را خروجی می‌دهد[۱۵]. به طور معمول، ستون فقرات یک شبکه

[1] Faster RCNN
[2] Region Proposal Network (RPN)
[3] Rol polling
[4] Pascal VOC
[5] MS COCO
[6] patch
[7] One-stage
[8] Feature map
[9] Single-shot detector (SSD)
[10] Retina Net
[11] CenterNet
[12] Two-stage
[13] Backbone architectures
[14] RGB

عصبی باقیمانده[1] مانند رِس–نِت–۵۰[2] است و قبل از تنظیم دقیق آن به وظایف پایین دستی، از قبل آموزش دیده است. شبکه عصبی باقیمانده[۱۶]، اولین شبکه عصبی پیشخور بسیار عمیق با صدها لایه، بسیار عمیق‌تر از شبکه‌های عصبی قبلی است که اتصالات پرش یا میانبرها برای پرش از روی برخی از لایه‌ها استفاده می‌شوند.

۳-ویژگی‌های تک مقایسی و چند مقیاسی[3]

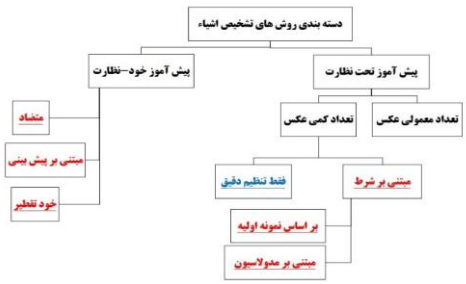
ابتدا به بیان مفهوم تانسور[4] پرداخته می‌شود. در ریاضیات، تانسور یک شیء جبری است که یک رابطه چند خطی بین مجموعه‌هایی از اشیاء جبری مرتبط با فضای برداری را توصیف می‌کند[۱۷]. اشیایی که تانسورها ممکن است بین آن‌ها ترسیم کنند شامل بردارها و اسکالرها و حتی تانسورهای دیگر هستند. ویژگی‌های تک مقیاسی، از یک تانسور سه بعدی تشکیل شده است که با گرفتن خروجی‌های یک لایه ستون فقرات خاص به دست می‌آید. ویژگی‌های چند مقیاسی از چندین تانسور سه بعدی در مقیاس‌های مختلف تشکیل شده است. صرفاً ترکیب چندین خروجی لایه از ستون فقرات باعث می‌شود که لایه‌های پایین‌تر با وضوح بالا اطلاعات معنایی محدودی داشته باشند.

۳-۱-مجموعه‌های داده[5]

معروف‌ترین مجموعه‌های داده، کلاس‌های شیء بصری پاسکال[6][۱۸] است که کلاس‌های شی در آن به ۲۰ دسته گسترش یافته است و اِم اِس کُوکُو[۱۹] که دارای ۹۱ شیء معمولی است که در بافت طبیعی آن‌ها یافت شده است که یک انسان چهار ساله به راحتی می تواند آن‌ها را تشخیص دهد. بیش از دو میلیون مورد و میانگین ۳.۵ دسته در هر تصویر دارد. علاوه بر این، شامل ۷.۷ نمونه در هر تصویر است که به راحتی بیشتر از سایر مجموعه داده های محبوب است و شامل تصاویر از دیدگاه های مختلف نیز می شود.

۳-۲-دسته‌بندی روش‌های تشخیص اشیاء

شکل ۳ دسته‌بندی روش‌های تشخیص اشیاء را نمایش می‌دهد:



شکل ۳ - دسته‌بندی روش‌های تشخیص اشیاء

همانطور که در شکل ۳ مشخص شده است، روش‌های مبتنی بر منطقه، یک خط در زیر آن‌ها کشیده شده است. همچنین روش‌های دو–مرحله‌ای با رنگ آبی و روش‌های تک–مرحله‌ای با رنگ قرمز مشخص شده است.

۳-۲-تعداد کمی عکس

تشخیص شء مبتنی بر تعداد کمی عکس، به تشخیص اشیاء تنها با استفاده از یک یا چند مثال آموزشی در هر کلاس می‌پردازد. این روش، اشیاء را به دو مجموعه مجزا از دسته‌ها تقسیم می‌کند: کلاس‌های پایه یا شناخته شده (کلاس‌های منبع)، که دسته‌های شیء هستند که برای آن‌ها به تعداد زیادی نمونه آموزشی در دسترس است و کلاس‌های جدید یا نادیده (کلاس‌های هدف) که برای هر کلاس فقط چند نمونه آموزشی (تصویر) موجود است. در اکثر این روش‌ها، فرض می‌شود که ستون فقرات آشکارساز شیء قبلاً روی یک مجموعه داده طبقه‌بندی تصویر مانند شبکه–تصویر (معمولا رِس–نِت–۵۰ یا ۱۰۱) از قبل آموزش داده شده است. روش‌های فقط تنظیم دقیق[7]، معمولاً از یک آشکارساز شیء سنتی مانند شبکه‌های عصبی پیچشی مبتنی بر منطقه سریع‌تر، با تغییرات جزئی معماری شروع می‌شوند و آموزش پایه را بر روی بسیاری از نمونه‌های کلاس پایه انجام می‌دهند، سپس بر روی یک مجموعه پشتیبانی که شامل کلاس‌های پایه و جدید است، تنظیم دقیق چند تصویر را انجام می‌دهند. منطق پشت این فرآیند دو مرحله‌ای، مقابله با عدم تعادل شدید بین کلاس‌های پایه و بدیع، و اجتناب از تطبیق بیش از حد در کلاس‌های بدیع است. روش‌های مبتنی بر شرط[8]، به دو دسته‌ی مبتنی بر نمونه اولیه[9] و مبتنی بر مدولاسیون[10] تقسیم‌بندی می‌شوند. در روش‌های مبتنی بر نمونه اولیه، نمونه‌های اولیه برای هر دسته از شیء، با شبکه هرمیِ دارای ویژگی تغییر شکل پذیر[11]، آموزش داده می‌شوند. در زمان آموزش پایه، چندین نمونه‌ی پشتیبان برای هر کلاس نمونه‌برداری می‌شود، نمایش‌های (نمایندگان) گردآوری شده، با ادغام منطقه مورد علاقه[12] آن‌ها را محاسبه می‌کند و پیشنهادات شیء را بر اساس فاصله آنها با نمایندگان طبقه‌بندی می‌کند. روش‌های مبتنی بر مدولاسیون معمولاً وزن‌های پشتیبانی (همچنین به عنوان وزن‌ها، نمونه‌های اولیه یا بردارهای توجه شناخته می‌شوند) را از ویژگی‌های پشتیبانی با استفاده از یک شاخه شرطی‌سازی جداگانه محاسبه می‌کنند.

[1] Residual neural network (ResNet)
[2] ResNet-50
[3] Single-scale and Multi-scale
[4] Tensor
[5] Data set
[6] Pascal Visual Object Classes (VOC)
[7] Fine tuning only
[8] Conditional based
[9] Prototype based
[10] Modulation based
[11] Deformable Feature Pyramid Network (FPN)
[12] Region of Interest (ROI) pooling

٤-پیش‌آموزِ خود-نظارت

پیش‌آموزشِ خود-نظارت، به عنوان یک جایگزین مؤثر برای پیش‌آموزشِ تحت-نظارت، ارائه شده است (نظارت از خود داده‌ها ناشی می‌شود). در این روش، به طور خودکار، برچسب‌ها را از داده‌های بدون برچسب، تولید می‌نمایند و برچسب‌ها، دوباره پیش بینی می‌شوند. رویکردهای پیش‌بینی‌کننده[1]، دی تِر[2] را با پیش‌بینی مجدد موقعیت که به‌طور خودکار تولید می‌شوند، از قبل آموزش می‌دهند. دی تِر، به تشخیص شیء، به عنوان یک مساله پیش‌بینی نگاه می‌کند و پیش‌بینی‌های منحصربه‌فردی را از طریق تطبیق دو بخشی و معماری رمزگذار-رمزگشای ترانسفورماتور[3] انجام می‌دهد. این محصولات یا به صورت تصادفی یا با استفاده از جستجوی انتخابی ، یک اکتشاف بدون آموزش مبتنی بر ادغام تکراری مناطق با رنگ‌ها، بافت‌ها و سایر ویژگی‌های محلی مشابه تولید می‌نمایند.بر خلاف روش‌های متضاد[4] عمومی که نمایش‌های کلی را در سطح تصویر متضاد می‌کنند، روش‌های متضاد محلی، نمایش‌های ستون فقرات را، چه در سطح ویژگی و چه در سطح برش، به صورت محلی، با امید به یادگیری بازنمایی‌های آگاه از مکان، متضاد می‌کنند.رویکردهای خود-تقطیر[5]، با به حداکثر رساندن شباهت بین پیش‌بینی‌های یک یاد دهنده و یک مدل یاد گیرنده، از یادگیری متضاد فاصله می‌گیرند.

٥-نتیجه‌گیری و بحث

تشخیص اشیا در دهه گذشته پیشرفت چشمگیری داشته است. این الگوریتم تقریباً در برخی از حوزه‌ها، به دقت سطح انسانی رسیده است، با این حال هنوز چالش‌های هیجان‌انگیزی برای مقابله با آن دارد. الگوریتم‌های تشخیص شیء، معمولاً از روش‌های یادگیری ماشین یا روش‌های یادگیری عمیق استفاده می‌نمایند. روش‌های تشخیص اشیاء می‌توانند تک-مرحله‌ای یا دو-مرحله‌ای باشند. هم زمان این روش‌ها، مبتنی بر منطقه، مبتنی برتبدیل کننده و مبتنی بر پیش-آموزش می-باشند. همچنین، این روش‌ها مبتنی بر پیش-آموزشِ تحت نظارت و یا پیش-آموزشِ خود-نظارت، هستند. علاوه بر این، ممکن است، مجموعه داده‌های از پیش برچسب‌گذاری‌شده، برای تشخیص اشیاء، وجود نداشته باشند. بنابراین، معیار استانداردی که این روش‌ها با آن مقایسه و ارزیابی می‌شوند، می‌تواند مبتنی بر تشخیص اشیاء معمولی، تشخیص اشیاء مبتنی بر تعداد کمی عکس و یا تشخیص اشیاء مبتنی بر تعداد اندک عکس، باشد. با توجه به این که شبکه‌های عصبی پیچشی مبتنی بر منطقه، خانواده‌ای از مدل‌های شبکه‌های عصبی پیچشی هستند که برای تشخیص اشیاء مورد استفاده قرار می‌گیرند، و دارای این خصوصیت می‌باشند که در حالتِ کاستی‌های طبقه‌بندی، علاوه بر استخراج ویژگی، اجزای بیشتری را نیز از قبل آموزش می‌دهند، با توجه به تعدد و هم‌پوشانی مقاله‌های روش‌های دسته بندی، در این مقاله، به دسته‌بندی روش‌های تشخیص اشیاء مبتنی بر منطقه، بطور جامع، پرداخته شد.

مراجع

[1] W.-Y. Lu and Y. Ming, "Face detection based on viola-jones algorithm applying composite features," in 2019 International Conference on Robots & Intelligent System (ICRIS), 2019: IEEE, pp. 82-85.

[2] A. Jain and D. Singh, "A Review on Histogram of Oriented Gradient," IITM Journal of Management and IT, vol. 10, no. 1, pp. 34-36, 2019.

[3] Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou, "A survey of convolutional neural networks: analysis, applications, and prospects," IEEE transactions on neural networks and learning systems, 2021.

[4] O. Russakovsky et al., "Imagenet large scale visual recognition challenge," International journal of computer vision, vol. 115, no. 3, pp. 211-252, 2015.

[5] W. Zhiqiang and L. Jun, "A review of object detection based on convolutional neural network," in 2017 36th Chinese control conference (CCC), 2017: IEEE, pp. 11104-11109.

[6] P. Bharati and A. Pramanik, "Deep learning techniques—R-CNN to mask R-CNN: a survey," Computational Intelligence in Pattern Recognition, pp. 657-668, 2020.

[7] R. Girshick, "Fast r-cnn," in Proceedings of the IEEE international conference on computer vision, 2015, pp. 1440-1448.

[8] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 779-788.

[9] H. Jiang and E. Learned-Miller, "Face detection with the faster R-CNN," in 2017 12th IEEE international conference on automatic face & gesture recognition (FG 2017), 2017: IEEE, pp. 650-657.

[10] X. Dong, L. Zheng, F. Ma, Y. Yang, and D. Meng, "Few-shot object detection," arXiv preprint Arxiv:1706.08249, pp. 1-11, 2017.

[11] T. Diwan, G. Anirudh, and J. V. Tembhurne, "Object detection using YOLO: challenges, architectural successors, datasets and applications," Multimedia Tools and Applications, pp. 1-33, 2022.

[12] W. Liu et al., "Ssd: Single shot multibox detector," in European conference on computer vision, 2016: Springer, pp. 21-37.

[13] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in Proceedings of the IEEE international conference on computer vision, 2017, pp. 2980-2988.

[14] X. Zhou, D. Wang, and P. Krähenbühl, "Objects as points," arXiv preprint arXiv:1904.07850, 2019.

[15] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 2117-2125.

[1] Prediction based
[2] DeTr
[3] Transformer encoder-decoder architecture
[4] Contrastive
[5] Self-distillation

[16] M. Thorpe and Y. van Gennip, "Deep limits of residual neural networks," arXiv preprint arXiv:1810.11741, 2018.
[17] W. H. Hampton, I. M. Hanik, and I. R. Olson, "Substance abuse and white matter: Findings, limitations, and future of diffusion tensor imaging research," Drug and alcohol dependence, vol. 197, pp. 288-298, 2019.
[18] D. Gudelj, A. Stama, J. Petrović, and P. Pale, "Visual Object Detection-an Overview of Algorithms and Results," in 2021 44th International Convention on Information, Communication and Electronic Technology (MIPRO), 2021: IEEE, pp. 1727-1732.
[19] T.-Y. Lin et al., "Microsoft coco: Common objects in context," in European conference on computer vision, 2014: Springer, pp. 740-755.