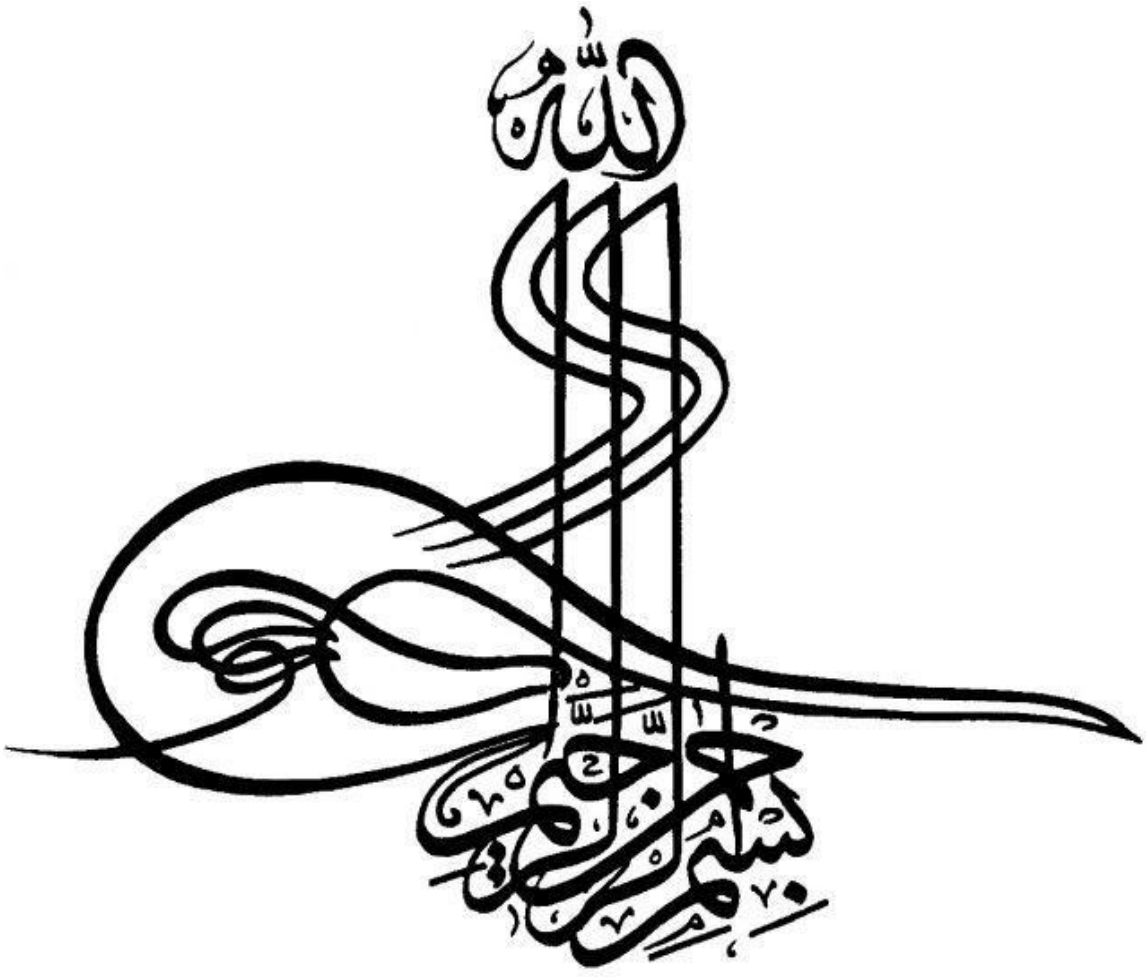

الله
بسم الله الرحمن الرحيم

دانشگاه خوارزمی تهران

دانشکده فنی و مهندسی

پایان‌نامه کارشناسی ارشد

رشته مهندسی کامپیوتر- گرایش هوش مصنوعی

عنوان:

بازشناسی صحنه با استفاده از ترکیب ویژگی‌های سراسری و محلی تصویر

نگارش:

فاطمه قنبری عدیوی

۹۱۳۸۲۱۵۲۵

استاد راهنما:

دکتر جمشید شنبه زاده

استاد مشاور:

دکتر زینب قصابی

بهمن ماه ۱۳۹۳

به نام خدا

دانشگاه خوارزمی تهران

دانشکده فنی و مهندسی

رساله کارشناسی ارشد

عنوان:

بازشناسی صحنه با استفاده از ترکیب ویژگی‌های سراسری و محلی تصویر

نگارش:

فاطمه قنبری عدیوی

کمیته ممتحنین:

استاد راهنما:

دکتر جمشید شنبه‌زاده **امضاء: .....**

استاد مشاور:

دکتر زینب قصابی **امضاء: .....**

استاد مدعو:

دکتر علی ذاکرالحسینی **امضاء: .....**

تاریخ: .....

تقدیم به همسر عزیزم

که سایه مهربانیش سایه سار زندگیم است،

او که اسوه صبر و تحمل بوده

و مشکلات مسیر را برایم آسان نمود.

با تقدیر و تشکر از اساتید ارجمندم

جناب آقای دکتر جمشید شنبه‌زاده و سرکار خانم دکتر زینب قصابی

که در کمال سعه صدر، با حسن خلق و فروتنی، از هیچ کمکی در این عرصه بر من دریغ ننمودند.

چکیده

بازشناسی صحنه یکی از مسائل چالش‌برانگیز در بینایی ماشین است که شامل دو مرحله استخراج ویژگی و انتخاب کلاس‌بند برای تصمیم‌گیری می‌باشد. استخراج ویژگی نقش کلیدی در بازشناسی صحنه دارد. روش‌های پیشنهادی برای بازشناسی صحنه را می‌توان از دو دیدگاه سطح و مقیاس استخراج ویژگی تقسیم‌بندی کرد. از دیدگاه سطح، یک تصویر را می‌توان با ویژگی‌های سطح پایین (مانند رنگ، بافت و لبه) یا سطح مفهومی (مانند جیست صحنه) بازنمایی کرد. از دیدگاه مقیاس، روش‌ها از مقیاس‌های مختلف محلی (بلوک‌ها، اشیاء، نواحی) و سراسری (کل تصویر) برای استخراج ویژگی استفاده می‌کنند. برخلاف این‌که برخی چارچوب‌های پیشنهادی، بازشناسی صحنه را با استفاده از ویژگی‌های سطح پایین انجام می‌دهند، شکاف معنایی میان این ویژگی‌ها و مفاهیم معنایی سطح بالا، پژوهشگران را مجبور به استفاده از ویژگی‌های سطح مفهومی همچون جیست کرده است. همچنین به دلیل عدم توجه به رابطه مکانی بین اجزای تصویر در روش‌های مبتنی بر ویژگی‌های سراسری، استفاده از ویژگی‌های محلی همراه با ویژگی‌های سراسری ضرورت می‌یابد.

در این پایان‌نامه یک رویکرد جدید روی مجموعه داده ۸ دسته صحنه خارجی (که شامل ۲۶۸۸ تصویر از دسته‌های ساحل، جنگل، بزرگراه، درون شهر، کوه، منظره باز، خیابان و ساختمان بلند است) براساس روش کلاس‌بندهای ترکیبی ارائه شده است. در روش پیشنهادی پس از به‌دست آوردن دو بردار ویژگی براساس ویژگی سراسری جیست و ویژگی محلی سیفت برای هر تصویر، دو کلاسبند ماشین بردار پشتیبان را به صورت جداگانه آموزش دادیم و نتایج خروجی را با استفاده از قانون ترکیبی ضرب برای به‌دست آوردن کلاس نهایی ترکیب کردیم.

با این روش توانستیم دقت کلاس‌بندی الگوریتم را برای این مجموعه داده با حفظ مزایای دو توصیفگر و بدون افزایش اندازه بردار ویژگی بهبود دهیم. نتایج عملی روی مجموعه داده موردنظر و مقایسه آن‌ها با برخی از روش‌های موجود با استفاده از معیارهای ارزیابی استاندارد نشان می‌دهند که روش پیشنهادی کارایی قابل قبولی در مقایسه با روش‌های قبلی به‌دست آورده است.

واژه‌های کلیدی: بازشناسی صحنه، جیست، سیفت، صحنه خارجی، ماشین بردار پشتیبان

فهرست مطالب

عنوان صفحه

فهرست شکل‌ها و جدول‌ها

فهرست علائم و اختصارات

HOG	Histogram of Gradiant
SVM	Support Vector Machine
LSVM	Linear Support Vector Machine
pLSA	Probabilistic latent Semantic Analysis
LDA	Latent Dirichlet Analysis
RSVP	Rapid Serial Visual Presentation
SUN dataset	Scene Understanding dataset
HIK	Histogram Intersection kernel
CMCT	Contextual Mean Census Transform
PCA	Principal Component Analysis

فصل اول: مقدمه و طرح تحقیق

۱-۱- مقدمه

یکی از مهم‌ترین اهداف هوش مصنوعی توسعه عوامل خودکاری است که می‌توانند محیط اطرافشان را از ورودی‌های بصری در حالت تصویر یا دنباله‌های ویدئویی در بسیاری کاربردها مانند جهت‌یابی ربات‌ها، دیدبانی محیط، تحلیل تصاویر پزشکی برای عمل‌های با مشارکت کامپیوتر و ادراک صحنه تفسیر کنند. هریک از این کاربردها نیازمند استدلال پیچیده درباره کنش‌های بین موجودیت‌های دنیای واقعی است. برای مثال، برای درک یک عکس از یک صحنه بصری نیاز به استدلال درباره اشیاء، نواحی پس‌زمینه، رنگ سطح، ساختار سه بعدی، مقیاس تصویر، نور، زاویه دید و کنش‌های پیچیده بین همه آن‌ها داریم. درمیان کاربردهای نامبرده، ادراک صحنه توجه خاصی از زمان تولد این موضوع در بینایی ماشین دریافت کرده است.

باتوجه به این که درک صحنه[۱] برای اکثر ناظران به راحتی صورت می‌گیرد، ممکن است تصور شود به همان آسانی درک می‌شود. بررسی دقیق‌تر نشان می‌دهد که ادراک صحنه یک فعالیت بسیار پیچیده است، که با چندین مسئله دشوار در ارتباط است: یک صحنه دقیقاً چیست؟ چه جنبه‌هایی از آن را به ما نشان می دهد؟ و چه فرآیندهایی در آن درگیر هستند؟ پیدا کردن پاسخ این پرسش‌ها دشوار است.

با این حال، پاسخ‌ها تا حدودی پیدا شده‌اند و یک درک کلی از ادراک صحنه آغاز شده است. جالب توجه است که این تصور ظاهرشده نشان می‌دهد که بسیاری از تجربه ذهنی ما به عنوان ناظران حداقل در مورد نحوه‌ی ادراک صحنه بسیار گمراه کننده است. به خصوص، این بازنمایی تصویری در ذهن ما تا حد زیادی یک توهم است.

۱-۲- تعریف صحنه

هندرسون[۲] و هولینگ ورث[۳] صحنه را یک دید مقیاس شده با انسان منسجم معنایی از محیط دنیای واقعی شامل عناصر پس‌زمینه و چندین شیء مجزا که با ترتیب مکانی چیده شده‌اند، تعریف می‌کنند. طبق گفته اولیوا[۴] صحنه جایی است که ما بتوانیم در آن حرکت کنیم. برای تمییز صحنه از «شیء» یا «بافت» فاصله بین بیننده و ناحیه ثابت را به عنوان فاکتور جداکننده در نظر می‌گیریم. بنابراین، یک «شیء» چیزی است که در محیط ۱ تا ۲ متری اطراف بیننده باشد اما برای یک صحنه فاصله بین بیننده و نقطه ثابت معمولاً بیش از ۵ متر است [۵].

---

[۱] scene understanding
[۲] Henderson
[۳] Hollingworth
[۴] Oliva

۱-۳- ادراک صحنه

۱-۳-۱- تعریف ادراک صحنه

ما در جهان صحنه‌هایی می بینیم، جایی که اشیای بصری که اغلب در یک مفهوم رایج با دیگر اشیای مرتبط تعبیه شده‌اند، قرار دارند. مغز انسان چگونه تحلیل می‌کند و این پیوستگی‌ها و مفاهیم مشترک را برای ادراک صحنه به کار می‌گیرد. توانایی مغز در تحلیل صحنه از این حقیقت می‌آید که جهان بصری اساساً تصادفی نیست و ساختار عناصر صحنه، پیکربندی طرح مکانی و ظاهر بصری در طول زمان به طور اساسی تغییر نمی‌کنند.

یک صحنه می‌تواند به صورت مجموعه‌ای از اشیا/ مفاهیم قرار گرفته در ترتیب های نسبتاً ثابت تعریف شود، اگرچه مسئله اصلی، نگرانی دانشمندان شناختی درباره این‌که «ماهیت بازنمایی‌های صحنه در مغز انسان در میان ویژگی‌های مکانی و مفهومی چیست؟» است، این پرسش پایه یک پاسخ مستقیم ندارد، اما ما می‌توانیم درباره این مسئله در یک سطح بسیار ساده از ادراک با مطرح کردن پرسش «آیا صحنه‌ها به عنوان مجموعه‌ای از اشیای مرتبط هم‌رخداد و معنایی (شیء- محور[۱]) درک می‌شوند یا فقط شکل کلی ساختار صحنه بدون توجه به اشیای موجود در آن (صحنه- محور[۲]) ؟». برای تشریح بیشتر، یک صحنه اداره را تصور کنید. یک اداره شامل اشیایی است که در دنیای واقعی هم‌رخداد هستند: صندلی‌ها، کامپیوترها، تلفن‌ها، خودکارها، کاغذها، کتاب‌ها و غیره. علاوه بر این اشیاء، اداره‌ها معمولاً یک ساختار صحنه مشخص دارند: چهار دیوار، کف، سقف، پنجره‌ها و شاید چند قفسه کتاب و سطوح میز مانند که به دیوار وصل شده‌اند. این ساختار صحنه، یک فضای سه‌بعدی با اشیایی که می‌توانند با روابط مکانی منسجم با یکدیگر قرار گیرند، را بازنمایی می‌کند. با این تفکیک، پژوهشگران علوم شناختی و روانشناسان، ایده‌های مختلفی در اهمیت اشیاء در مقابل طرح کلی صحنه دارند که می‌توانند به دو رویکرد متفاوت دسته‌بندی شوند: رویکردهای شی- محور و رویکردهای صحنه- محور.

۱-۳-۲- مسئله ادراک صحنه

ادراک صحنه فرآیند تبدیل بازنمایی‌های دوبعدی تصویر به شکلی نمادین از دانش برای توصیف تصویر با اشیای درون آن (مانند افراد، خیابان، کتاب، دریا و غیره)، محیط صحنه‌ای که تصویر می‌شود (مانند داخلی[۳]، اداره، غروب آفتاب، شهر و غیره)، رابطه بین عناصر صحنه (عناصر پیش‌زمینه و پس‌زمینه) و

---

[۱] object-centered
[۲] scene-centered
[۳] indoor

فعالیت/ رخدادی که شامل می‌شود (مانند تنیس بازی کردن، شنا کردن و غیره) است. به عبارت دیگر هدف از ادراک صحنه پاسخ به پرسش‌های چه چیزی (برچسب فعالیت/ رخداد)، کجا (برچسب صحنه)، چه کسی (اشیای در صحنه) و چگونه (رابطه مکانی، پیکربندی صحنه و طرح مکانی) برای تصویر داده شده، است. یک تصویر صحنه در شکل ۱-۱ نشان داده شده است. یک سیستم درک صحنه این مفهوم معنایی را به عنوان مجموعه‌ای از اشیاء مانند کوه، آب، افراد و آسمان (چه کسی)، قایقرانی به عنوان رخداد (چه چیزی)، کوه در بالای دریا واقع شده، آسمان با کوه پوشانده شده (چگونه) و خارجی[۱]/ اقیانوس/ طبیعت به عنوان برچسب محیط (کجا) تفسیر می‌کند. اگرچه، سیستم ادراک صحنه نمی‌تواند بعضی پرسش‌های مرتبط با این تصویر صحنه مثل مرد چه کسی است؟ (مسئله تعیین هویت)، آن مکان کجاست؟ (ساحل غربی یا شرقی هاوایی[۲])، احساس مرد چیست؟ (آشکارسازی حالت/ احساس چهره)، زمان و تاریخ دقیقی که این تصویر گرفته شده، چه موقع است؟ (جمعه ۲۱ام ژوئن یا دوشنبه ۲۳ام ژوئن) را پاسخ دهد.

شکل ۱-۱ نمونه‌ای از ادراک صحنه؛ اشیاء، کوه، آب، اشخاص و آسمان (چه کسی)، فعالیت: قایقرانی (چه چیزی)، رابطه عناصر: کوه بالای آسمان است، آسمان با کوه هم‌پوشانی دارد (چگونه) و برچسب صحنه: خارجی/ اقیانوس/ طبیعت (کجا) [۴۷].

[۱] outdoor
[۲] Hawaii

اگرچه ادراک صحنه در تعریف، ساده است، مدل کردن یک سیستم خودکار برای تحلیل یک تصویر صحنه یک مسئله چالش‌برانگیز در بینایی ماشین است. با مراجعه به برخی رویکردهای بینایی محاسباتی، بسیاری از پژوهشگران، چالش ادراک صحنه و برخی سیستم‌های موثر که ساخته شده را برعهده گرفتند. اگرچه، این سیستم‌های اولیه، در مقیاس‌گذاری مسائل دنیای واقعی بزرگ شکست خورده‌اند پژوهشگران را مجبور به تغییر جهت به جنبه‌های اختصاصی مسئله ادراک صحنه در جداسازی، همچون آشکارسازی شیء، قطعه قطعه‌سازی تصویر و بازسازی سه بعدی شده‌اند [۲۴].

۱-۴- سیستم‌های ادراک صحنه

۱-۴-۱- تعریف سیستم‌های ادراک صحنه

سیستم‌های ادراک صحنه براساس دو زمینه مطالعه بزرگ ساخته شده‌اند: یادگیری ماشین و بینایی محاسباتی. کاربردهای بینایی محاسباتی سطح بالا هم‌اکنون با روش‌های یادگیری ماشین تا حد زیادی گره خورده‌اند. همچنین، بسیاری از روش‌های یادگیری ماشین الهام گرفته از مسائل ناشی از بینایی محاسباتی هستند و بنابراین گسترده‌تر اعمال شده‌اند. علاوه بر روش‌های محاسباتی مطرح شده برای ادراک صحنه، بسیاری مطالعات پژوهشی در روانشناسی و علوم شناختی برای یافتن این‌که مغز انسان چگونه صحنه‌های اطرافش را درک می‌کند، انجام می‌شوند.

۱-۴-۲- کلاس‌بندی[۱] صحنه توسط ماشین

درمیان همه کارهای مرتبط با ادراک صحنه، کلاس‌بندی صحنه مفیدترین اطلاعات سطح بالا را درباره محیط صحنه، برچسب صحنه، این‌که چه چیزی را به تصویر می‌کشد (داخلی در مقابل خارجی، طبیعی[۲] در مقابل مصنوعی[۳] ، اداره، خیابان، ساحل و غیره) بیان می‌کند. علاوه بر این، کلاس‌بندی صحنه می‌تواند برای آسان کردن دیگر کارهای ادراک صحنه مانند بازشناسی شیء، قطعه قطعه‌بندی، تفسیر سه بعدی و تخمین زاویه دید به کار رود. برای مثال، دانستن این‌که یک صحنه، ساحل است، این محدودیت را به وجود می‌آورد که در صحنه باید شن و ماسه وجود داشته باشد. شکل ۱-۲ برخی نمونه‌های صحنه را به صورت دسته‌های طبیعی/ مصنوعی، داخلی/ خارجی، اداره، باغ‌وحش، منظره طبیعی و شهر کلاس‌بندی کرده است. یک سیستم ساده کلاس‌بندی صحنه می‌تواند به صورت دو واحد[۴] درنظر گرفته شود: واحدهای استخراج

---

۱ classification
۲ natural
۳ man-made
۴ module

ویژگی[1] و کلاس‌بندی صحنه. واحد استخراج ویژگی تلاش می‌کند جنبه‌های بصری صحنه را از تصویر صحنه دوبعدی که روش‌های یادگیری بر آن تکیه می‌کنند، مشخص کند. هنگامی که یک تصویر صحنه خام- که با یک شبکه دوبعدی از مقادیر پیکسل گسسته توصیف می‌شود- با درک انسان قابل‌فهم باشد، به اندازه کافی نمایشگر ادراک توسط ماشین به دلیل توصیف محدود ویژگی‌های اولیه تصویر و پرمایگی معناهای انسانی نیست، که به عنوان «شکاف معنایی[2]» نام برده می‌شود. برای دنبال کردن مسئله شکاف معنایی یک توصیف جامع درباره اشیایی که صحنه شامل می‌شود (مانند اتومبیل‌ها و افراد)، طرح مکانی که اشیا جای گرفته‌اند و رابطه‌ی مفهومی آن‌ها (به عنوان مثال، یک موشواره روی یک میز)، یا یک کلاس که صحنه متعلق به آن است (مانند یک صحنه شهری) برای یادگیری نیاز است. این سطوح، اطلاعات معنایی مختلفی درباره یک صحنه به‌دست آورده و رویکردهای محاسباتی مختلف برای استخراج این اطلاعات نیاز دارد. بنابراین، واحد کلاس‌بندی تلاش می‌کند بازنمایی قابل یادگیری از تصویر با روش‌های یادگیری ماشین و استدلال را برای انجام کلاس‌بندی مفهوم صحنه به کار گیرد.

شکل ۱-۲ نمونه‌های کلاس‌بندی صحنه، الف- مصنوعی/ خارجی/ شهر، ب- مصنوعی/ داخلی/ اداره، ج- طبیعی/ خارجی/ باغ‌وحش، د- طبیعی/خارجی/فضای باز.

[1] feature extraction
[2] semantic gap

۱-۵- جایگاه و اهمیت موضوع

برای مشخص کردن رویکرد بازشناسی، فعالیت‌های پژوهشی با ۵ محور زیر سازماندهی شده است. برای هر محور چالش‌های علمی اصلی آن خلاصه شده است:

۱- ادراک برای درک صحنه (دنیای ادراکی): اولین جمع‌آوری و توسعه الگوریتم‌های بینایی برای کار با شرایط متفاوت دنیای واقعی است. هدف این الگوریتم‌ها آشکارسازی و کلاس‌بندی اشیای فیزیکی موردنظر است. رایج‌ترین الگوریتم‌ها حرکت ویدئوها را تخمین می‌زند. این الگوریتم‌ها براساس این فرض است که اشیای مورد نظر[۱] مرتبط با چیزی است که در ویدئو حرکت می‌کند، که می‌تواند برای آشکارسازی تغییرات سیگنال استنتاج کند. متاسفانه این الگوریتم‌ها گرایش به آشکارسازی پارازیت زیادی (بخاطر تغییرات نور) با دیگر اشیای موردنظر را دارند. مشکل دوم استخراج ویژگی‌های مشخصه اشیای موردنظر است. اغلب الگوریتم‌ها ویژگی‌های مرتبط با خط سیر اشیای فیزیکی را محاسبه می‌کنند.

توصیفگرهای قوی که مشخصه‌ی شکل اشیای فیزیکی هستند هنوز نیاز به توسعه دارند. سومین مشکل تولید فرضیه‌هایی با الگوریتم‌های معتبر و درک محدودیت‌های آن‌هاست. با روشی مشابه، الگوریتم‌ها، دیگر رسانه‌ها و شرایط (مانند: صدا، مخاطب، رادار) را که نیاز به توسعه بیشتر برای تکمیل اطلاعات استخراج شده از جریان‌های ویدئو دارد، پردازش می‌کند. یک مسئله که هنوز حل نشده برقراری درک و احتمال این فرایندهاست.

۲- نگهداری وابستگی سه بعدی در طول زمان (دنیای واقعی): دومین محور شامل ترکیب همه اطلاعات به‌دست آمده از حسگرهای مختلف که در حین آشکارسازی اشیای فیزیکی و دنبال کردن این اشیا در طول زمان مشاهده شده است. برخلاف همه کارهای انجام شده در این زمینه در طول ۲۰ سال اخیر الگوریتم‌های ترکیب و تعقیب[۲]، شکننده باقی می‌مانند. برای تضمین وابستگی این اشیای تعقیبی، دلیل مکان-زمانی[۳] نیاز است. یک مسئله باز دیگر مدل کردن عدم قطعیت این فرایندهاست. پرسشی دیگری که نیاز به پاسخگویی دارد این است که در چه سطحی این اطلاعات باید ترکیب شوند. ترکیب اطلاعات در سطح سیگنال اطلاعات دقیق‌تری را تولید می‌کند اما تعقیب اطلاعات در سطحی بالاتر و برای درک آسان‌تر است.

---

[۱] interest
[۲] fusion and tracking
[۳] spatio-temporal

در هر موردی، یک فرموله‌سازی دقیق برای ترکیب اطلاعات غیردقیق به‌دست آمده از حسگرهای ناهمگن (غیریکنواخت) نیاز است.

۳- آشکارسازی رویداد (دنیای معنایی): در سطح رویداد، پردازش روابط بین اشیای فیزیکی، محور سوم را تشکیل می‌دهد. چالش واقعی، بررسی موثر همه روابط مکانی- زمانی ممکن اشیایی است که ممکن است با رویدادها منطبق باشند؛ رویدادها با عمل‌ها[۱]، وضعیت‌ها[۲]، فعالیت‌ها[۳]، رفتارها[۴]، سناریوها[۵]، سندها[۶] و تاریخچه‌ها[۷] نام‌گذاری می‌شوند. تنوع این رویدادها که عموماً رویدادهای ویدئویی نامیده می‌شوند گسترده و وابسته به دانه های[۸] مکانی و زمانی آن‌ها با توجه به تعداد اشیای فیزیکی درگیر در رویدادها و پیچیدگی رویداد (تعداد مولفه‌های تشکیل‌دهنده رویداد و نوع رابطه زمانی) می‌باشد. بنابراین چالش، بررسی این فضای رویداد بزرگ بدون از دست دادن جستجوهای ترکیبی است.

۴- ارزیابی، کنترل و یادگیری (سیستم‌های خودمختار): برای توانایی بهبود سیستم‌های درک صحنه، از یک جنبه نیاز به ارزیابی کارایی آن‌ها داریم. روش کلاسیک برای ارزیابی کارایی شامل استفاده از داده مرجع (که حقیقت پایه[۹] نامیده می‌شود) است. اگرچه تولید حقیقت پایه طاقت‌فرسا و مستعد خطا است. بنابراین یک مسئله، اجرای مرحله ارزیابی کارایی با استفاده از روش‌های بدون ناظر است. اگر ارزیابی ممکن باشد، یک چالش واقعی بهینه‌سازی سیستم درک صحنه با استفاده از ارزش‌های یادگیری ماشین برای یافتن بهترین ترکیب برنامه‌ها، بهترین مجموعه از پارامترها با بهترین روش‌های کنترلی برای به‌دست آوردن یک فرآیند بلادرنگ موثر و کارا می‌باشد. برنامه‌ها وابسته به شرایط محیطی هستند و فرایند بهینه‌سازی برنامه باید با توجه به تغییرات محیط و منابع در دسترس، پویا باشند. مسئله دوم، همه این برنامه‌ها با یکدیگر در ارتباطند، بنابراین تغییر یک پارامتر برنامه می‌تواند عمل دیگر برنامه‌ها را تغذیه کند. درنهایت، دانش این برنامه‌ها فرموله نیست و معمولاً حتی توسعه‌دهندگان نمی‌توانند بگویند خروجی برنامه تحت شرایط خاصی چه خواهد بود. از طرف دیگر، برای بهبود کارایی سیستم اضافه کردن استدلال با اهمیت‌تر است. درک صحنه، یک رویکرد پایین به بالا شامل خلاصه کردن اطلاعات به‌دست آمده از سیگنال است (رویکرد با داده

---

۱ actions
۲ situations
۳ activities
٤ behaviors
٥ senarios
٦ scripts
۷ chronicles
۸ granularing
۹ ground truth

راهنمایی می‌شود). اگرچه، در بعضی موارد یک رویکرد بالا به پایین (رویکرد با مدل‌ها راهنمایی می‌شود) می‌تواند کارایی فرایند کمتر را با تهیه یک دانش کلی‌تر از صحنه مشاهده شده یا با بهینه‌سازی منابع در دسترس بهبود بخشد. برای مثال، وابستگی کلی دنیای چهاربعدی می‌تواند به تصمیم‌گیری در این‌که آیا بعضی نواحی متحرک منطبق با پارازیت یا اشیای فیزیکی موردنظر هستند، کمک کند.

بنابراین، محور چهارم در بررسی نظارت بر برنامه (شامل ارزیابی) و روش‌های یادگیری ماشین برای نسلی آسان از سیستم‌های درک صحنه بلادرنگ موثر است.

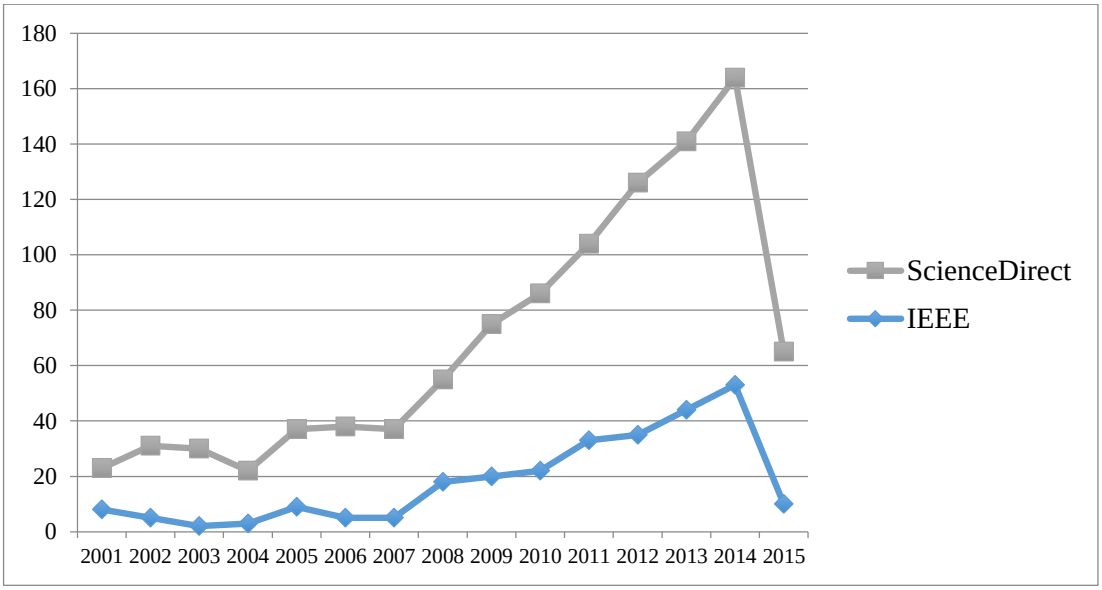
۵- ارتباط، تصویرسازی و جمع‌آوری دانش (سیستم تعاملی): حتی زمانی که تفسیر صحیحی از صحنه انجام شود، سیستم درک صحنه هنوز باید با درک آن با کاربران در ارتباط باشد. بنابراین تعاملات کاربران محور پنجم را تشکیل می‌دهد. حداقل سه نوع کاربر وجود دارد: توسعه‌دهندگان برنامه، متخصص حوزه کاربرد و کاربران نهایی. اولین چالش، قادر ساختن توسعه‌دهندگان برنامه برای درک همه مولفه‌های خاص و در زمانی یکسان، ساختار کلی سیستم درک صحنه است، بنابراین آن‌ها می‌توانند برنامه‌هایشان را بطور موثر سازگار سیستم را روی یک سایت پیکربندی و نصب کنند. برای رسیدن به این هدف، فرموله کردن برای بیان دانش برنامه نیاز است. دوماً اگر ما یک سیستم موثر بخواهیم؛ اطلاعات اولیه نیاز به فرموله شدن برای توانایی متخصص حوزه برای توصیف روش آن‌ها برای تحلیل صحنه است. برای مثال، یک ابزار و یک آنتولوژی اختصاصی باید برای کمک به متخصصان جهت دریافتن سناریوهایی که سیستم باید آشکارسازی کند، تهیه شود. برای کمک به این فرایند، یک ابزار گرافیکی می‌تواند برای تولید تصویرسازی انیمشن‌های مجازی سه بعدی توصیف این سناریوها طراحی شود. برای تکمیل این ابزارها، روش‌های دسته‌بندی می‌توانند برای استخراج فعالیت‌های تکراری رخ داده در صحنه به کار روند (الگوهای رویداد یا سری‌های زمانی). به علاوه، اگر ما بخواهیم از سیستم استفاده کنیم مراقبت خاصی برای نمایش آنچه کاربران نهایی باید درک کنند، نیاز است. یک واسط آرگونومیک روی یک رسانه منطبق (مثلاً دستیار دیجیتال شخصی[۱])، نمایشی ساده از صحنه (مثلاً صحنه سه بعدی مجازی) و یک واژه‌نامه شهودی وسایل مورد نیاز برای ارائه به کاربران نهایی هستند. علاوه بر این وسایل، سیستم نیاز به محاسبه اطلاعات بازخوردی از کاربران نهایی برای سازگاری کارایی آن با اهداف خاصی هر کاربر دارد.

روند رشد تحقیقات در زمینه بازشناسی صحنه در شکل ۱-۳ نشان داده شده و همان‌طور که مشخص است تعداد مقالات مرتبط، نشان‌دهنده این مطلب است با گذشت زمان، مرتباً بر اهمیت این مبحث افزوده

---

[۱] PDA

شده است. نمودار زیر بیان می‌کند که بازشناسی صحنه کماکان یکی از دغدغه‌های اصلی در زمینه یادگیری ماشین بوده و هر روزه تلاش‌های محققان برای بهبود آن بیشتر می‌شود.



شکل ۱-۳ نمودار رشد مقالات در زمینه بازشناسی صحنه درSienceDirect و IEEE.

۱-۶- بیان مسئله

برخلاف تعداد کمی سناریو موفق، مانند کنترل ترافیک و آشکارسازی نفوذ، سیستم‌های درک صحنه حساس باقی می‌مانند و می‌توانند فقط تحت شرایط محدود (مثلاً در طول روز نسبت به شب، شرایط روشنایی پراکنده و بدون سایه) در طول زمان با داشتن کارایی ضعیف، قابلیت اصلاح سخت، مقدار کم دانش اولیه نسبت به محیط عمل کنند. اگرچه این سیستم‌ها بسیار خاص هستند و نیاز به توسعه برای حذف کاربردهای دیگر دارند. برای پاسخ‌گویی به این مسائل، اغلب پژوهشگران برای توسعه الگوریتم‌های بینایی جدید با عملکردهای متمرکز و قدرت کافی برای کار با شرایط زندگی واقعی تلاش کرده‌اند. تا به امروز هیچ الگوریتم بینایی نتوانسته تغییرات زیاد شرایط مانند شرایط حسگرها، نیازمندی‌های سخت‌افزار، شرایط نوری، تغییرات فیزیکی شی، و هدف‌های کاربردی که مشخصه صحنه‌های دنیای واقعی است را نشانه‌گذاری کند. هدف، طراحی یک قالب برای تولید آسان سیستم‌های درک صحنه خودمختار و موثر است. این هدف

ایده‌ال است، اگرچه برخی از بهترین[۱] روش‌های امروزی در بینایی شناختی، منجر به روش‌های نیمه‌کاملی[۲] شده است. برای رسیدن به این هدف، یک رویکرد کلی‌نگر نیاز است که فرآیند درک صحنه اصلی براساس نگهداری وابستگی نمایش صحنه سه‌بعدی کلی در تمام زمان است. این رویکرد را می‌توان نمایش معنایی چهاربعدی نامید که براساس مدل‌ها و ثوابت، مشخصه صحنه و پویایی آن است. درک صحنه یک فرایند پیچیده است که اطلاعات را در چهار سطح خلاصه می‌کند. سیگنال (مانند پیکسل، صدا)، ویژگی‌های ادراکی[۳]، اشیای فیزیکی و رویدادها[۴]. سطح سیگنال با پارازیت[۵] قوی، مبهم، خراب و داده ازدست رفته، مشخص می‌شود.

فرایند کلی درک صحنه شامل فیلتر کردن این اطلاعات برای آوردن چهارمین بینش مرتبط صحنه و پویایی آن است. برای اجرای این هدف، مدل‌ها و ثوابت، نقاطی حیاتی برای مشخص کردن دانش و اطمینان از سازگاری (پایداری) آن در چهار سطح خلاصه شده‌اند. برای نمونه، تعریف فرمول‌هایی برای مدل کردن صحنه خالی اطراف (مثلاً هندسه آن)، حسگرها ماتریس‌های دوربین‌ها، اشیای فیزیکی مورد انتظار در صحنه (مثلاً مدل سه بعدی انسان بودن) و سناریوهای موردنظر کاربران (مانند رویدادهای غیرعادی)، ثوابت (قاعده[۶] نامیده می‌شوند) قوانین کلی مشخصه پویایی صحنه هستند. برای نمونه چگالی یک پیکسل می‌تواند فقط در دو حالت تغییر شرایط نوری (مانند سایه) یا تغییر به دلیل یک شی فیزیکی (مثلاً هم‌پوشانی) تغییر کند. قانون دوم، بررسی اشیای فیزیکی که نمی‌توانند در میانه صحنه ناپدید شوند. هنوز یک مسئله باز شامل تعیین این‌که این مدل‌ها ثوابت به عنوان یک پیشین داده شده یا آموزش داده شده است وجود دارد. چالش کلی، شامل سازماندهی همه این دانش‌ها به منظور کسب تجربه، اشتراک آن با دیگران و به‌روزرسانی آن در میان آزمایش‌ها است. برای رویایی با این چالش، ابزارهایی در مهندسی دانش مثل آنتولوژی نیاز می‌باشد.

در مسئله بازشناسی صحنه، استخراج ویژگی نقشی اساسی در عملکرد الگوریتم دارد. با شروع کار سیستم‌های بازشناسی صحنه، این مسئله مطرح شد که از چه ویژگی‌های از تصویر استفاده شود که بتوان میان دقت روش و زمان اجرای آن مصالحه‌ای مناسب برقرار کرد. به علاوه، روش پیشنهادی تا حد امکان نسبت به تغییراتی همچون تغییرات روشنایی، مقیاس، زاویه دید و وجود هم‌پوشانی، دگردیسی و پارازیت

---

۱ state-of-the-art
۲ partial
۳ perceptual features
۴ events
۵ noise
۶ regularity

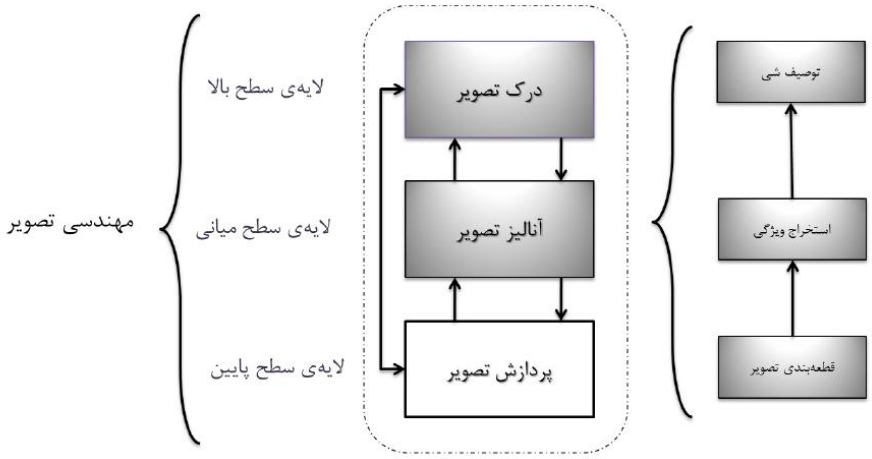
پایدار[1] باشد. مسئله اصلی این پایان نامه هم استخراج ویژگی‌های موثر سراسری و محلی از تصاویر صحنه برای ایجاد بردار ویژگی با ابعاد پایین‌تر از برخی روش‌های موجود که تا حد امکان نسبت به عوامل تغییر روشنایی، مقیاس، زاویه دید و تغییرات درون کلاسی پایدار باشند و بتوان با این ویژگی‌ها دقت و زمان بازشناسی صحنه را بهبود بخشید.

۱-۷- حوزه مسئله

یکی از مراحل مهم در انجام پایان‌نامه مشخص کردن حوزه مساله است تا از انجام فعالیت‌های غیرمرتبط با مساله جلوگیری شود، به همین دلیل در این بخش توضیح داده می‌شود که تحقیقات این پایان‌نامه در کدام حوزه‌ها قرار می‌گیرد.

فرآیند مهندسی تصویر در شکل ۱-۴ نشان داده شده است. در این مراحل ابتدا در سطح پایین تصویر پردازش می‌شود و پارازیت و مشکلات احتمالی تصویر برطرف می‌شود. این مولفه در حوزه تحقیق قرار ندارد و تصویر ورودی تصویری آماده و پردازش شده است. اما دو مولفه دیگر شامل آنالیز تصویر و درک تصویر در حوزه مساله قرار می‌گیرد و برای بازشناسی شیء آنالیز و درک تصویر انجام می‌شود. در مسئله بازشناسی صحنه معمولا از مجموعه داده‌های آماده استفاده می‌شود و این سخن به آن معنی است که بخش پردازش تصاویر اولیه در حیطه‌ی این موضوع نمی‌گنجد و این بخش در تمام روش‌های موجود یکسان در نظر گرفته می‌شود. بنابراین حوزه‌ی بازشناسی صحنه در دو حیطه‌ی آنالیز تصویر و درک تصویر گسترده شده است. بنابراین الگوریتم‌های بازشناسی صحنه بر روی بخش‌های قطعه‌بندی (در صورت نیاز)، استخراج ویژگی، توصیف این ویژگی‌ها و درک تصویر فعالیت دارند.

[1] invariant



شکل ۱-۴ فرآیند مهندسی تصویر که شامل سه مرحله پردازش، آنالیز و درک تصویر است [۴۳].

۱-۸- جمع‌بندی

این پایان‌نامه در شش فصل تنظیم شده است. در فصل ۲ ابتدا به معرفی ویژگی‌های اصلی که معمولاً در دسته‌بندی تصاویر استفاده می‌شود می‌پردازیم. همچنین مقالات مطرح در این زمینه معرفی شده و نقاط قوت و ضعف آن‌ها بیان می‌شود.

در فصل ۳ تئوری موردنیاز برای حل مسئله مطرح شده را با ذکر جزئیات تشریح می‌کنیم. همان‌طور که در ادامه می‌بینیم، در روش پیشنهادی از دو روش جیست نرمال شده[۱] و سیفت[۲] استفاده می‌شود که این دو روش در این فصل به تفصیل آمده‌اند.

در فصل ۴ مراحل الگوریتم پیشنهادی و جزئیات آن به تفصیل ذکر می‌شوند. همچنین در این فصل توضیحات مربوط به هر مرحله و تنظیم پارامترهای مربوط به آن‌ها آورده شده است.

در فصل ۵ نتایج انجام آزمایش بر روی مجموعه داده‌ی ۸ دسته صحنه خارجی که یکی از مشهورترین مجموعه داده‌های صحنه‌های خارجی است و توسط اولیوا و تورالبا[۳] جمع‌آوری شده است، آورده می‌شود. با استفاده از معیارهای ارزیابی استاندارد این نتایج با برخی از روش‌های مطرح قبلی و همچنین دو روش جیست نرمال‌شده و سیفت مقایسه شده است.

در فصل ۶ نتیجه‌گیری کلی کارهای انجام شده و پیشنهاداتی برای ادامه کار و بهبود روش ارائه می‌شود.

---

۱ normalizedgist
۲ scale invariant feature transform (sift)
۳ Torralba

فصل دوم : مروری بر کارهای انجام‌شده

۲-۱- مقدمه

در این فصل برخی از ویژگی‌هایی را که الگوریتم‌های دسته‌بندی تصاویر براساس آن‌ها کار می‌کنند بررسی می‌کنیم. یکی از مهم‌ترین چالش‌های این مبحث استفاده از ویژگی‌های مناسب جهت دسته‌بندی تصاویر است که در برابر تغییرات تصاویر مقاوم باشند. این امر توجه محققین را در دهه‌های اخیر به خود جلب کرده است، به گونه‌ای که هر ساله روش‌های متعددی در جهت بهبود این مبحث ارائه می‌شود.

همچنین برخی از الگوریتم‌هایی را که جهت استخراج ویژگی در پردازش تصویر طی چند سال اخیر از آن‌ها استفاده شده است شرح دادیم، حال در این بخش به معرفی جزئی‌تر چند نمونه از این ویژگی‌های پایه‌ای تصویر که نقش به‌سزایی در پیشبرد بازشناسی تصویر داشته‌اند، می‌پردازیم.

۲-۲- تاریخچه

اولین فعالیت‌ها در زمین درک دنیای مجازی در اواخر دهه ۸۰ و اوایل دهه ۹۰ چاپ شده است. این فعالیت‌ها کارهایی بسیار اولیه بودند؛ زیرا انجمن گرافیک کامپیوتری متقاعد نشده است که این زمینه برای گرافیک کامپیوتری مهم بوده است. هدف این فعالیت‌ها پیشنهاد به کاربر برای کمک به درک آسان دنیای مجازی با محاسبه زاویه دید خوبی است.

در کاربردهای بینایی محاسباتی، یک تصویر ورودی می‌تواند با ویژگی‌های سطح پایین اولیه مانند مقادیر پیکسل گسسته در فضای رنگ، بافت‌ها[۱] و نگاشت‌های لبه[۲] مختلف، که با ماشین قابل درک نیست نمایش داده شود. به عبارت دیگر، نبود بازنمایی سطح بالای یک تصویر ورودی، توانایی ماشین را برای تحلیل صحنه محدود می‌کند. به دنبال آن، روش‌های یادگیری ماشین برای حذف شکاف معنایی بین اطلاعاتی که قابل بازنمایی با ویژگی‌های سطح پایین هستند و توانایی ماشین برای تفسیر معنای سطح بالای صحنه، نیاز است. ابزار یادگیری ماشین رسمی به صورت روش‌های باناظر[۳] و بدون ناظر[۴] دسته‌بندی می‌شوند. هدف روش‌های باناظر پیش‌بینی مقدار یک اندازه‌گیری خروجی (مانند برچسب دسته) براساس مجموعه‌ای از ویژگی‌های ورودی و برچسب‌های داده است. درحالیکه، در یادگیری بدون ناظر، هیچ اندازه‌گیری خروجی وجود ندارد و هدف توصیف این است که چگونه داده ورودی برای به اشتراک گذاشتن ویژگی‌های مشترک برای توصیف هر زیرمجموعه مشابهی از داده ورودی، سازماندهی یا خوشه‌بندی شده است.

---

۱ texture
۲ edge
۳ supervised
۴ unsupervised

بینایی محاسباتی یک موضوع بسیار گسترده با زیربخش‌های بسیار است. در این بخش، مرور مختصری بر سطوح مختلف بازنمایی ویژگی در کلاس‌بندی صحنه انجام شده است. به علاوه، دو کار اساسی در بینایی ماشین با رویکردهای کلاس‌بندی صحنه همچون آشکارسازی شیء و قطعه قطعه‌بندی تصویر چندسطحی، ترکیب شده است.

۲-۳- ویژگی‌های تصویر

در کلاس‌بندی صحنه یک تصویر می‌تواند با ویژگی‌های سطح پایین (مانند رنگ، بافت و لبه) یا سطح مفهومی (مانند هم- رخدادی[۱] اشیاء یا طرح سه بعدی اشیای پیش‌زمینه روی سطح پس‌زمینه) بازنمایی شود. هنگامی‌که چارچوب‌های مطرح شده ساده‌تر، کلاس‌بندی صحنه را با استفاده از ویژگی‌های سطح پایین (مانند رنگ و بافت) اجرا کنند، شکاف معنایی ناشی از ویژگی‌های سطح پایین، پژوهشگران را مجبور به استفاده از ویژگی‌های سطح معنایی برای بهبود دقت کلاس‌بندی می‌کند.

۲-۳-۱- ویژگی‌های سطح پایین

ویژگی‌های سطح پایین به ویژگی‌های اولیه‌ای مانند رنگ، لبه، شکل و بافت گفته می‌شود که قابل استخراج از تصویر خام دوبعدی در گام آغازی پردازش تصویر باشند.

۲-۳-۱-۱- رنگ

یکی از پراستفاده‌ترین ویژگی‌ها در کاربردهای بینایی ماشین رنگ است. رنگ‌ها در یک فضای رنگی انتخابی به صورت RGB، LAB، LUV، HSV (HSL)، YCbCr و اختلاف فام کمینه- بیشینه (HMMD) تعریف می‌شوند. بازنمایی یک تصویر با یک آرایه دو بعدی از سه مولفه رنگی یا مقادیر خاکستری برای بازنمایی بصری و ذخیره‌سازی مفید است، اما خیلی برای ادراک خودکار صحنه مفید نیست. بنابراین، گام‌های اضافی برای استخراج ویژگی‌های توصیفی بیشتر از مولفه‌های رنگی پایه همچون ماتریس رنگ- کواریانس، هیستوگرام رنگ، گشتاورهای رنگ و بردار انسجام رنگ نیاز است.

۲-۳-۱-۲- بافت

اگرچه بافت، مانند ویژگی‌های رنگی خوش- تعریف نیست، اطلاعات مهمی در کلاس‌بندی صحنه فراهم می‌کند. بافت مفهوم تصاویر طبیعی بسیاری همچون پوست میوه، ابرها، درختان، خشت‌ها و تاروپود را توصیف می‌کند. بنابراین، بافت یک ویژگی مهم در تعریف معنا برای هدف بازیابی تصویر است.

---

[۱] co-occurrence

ویژگی‌های بافتی معمولاً در سیستم‌های کلاس‌بندی صحنه شامل ویژگی‌های طیفی، مانند ویژگی‌های به‌دست آمده از فیلترهای گابور[۱] یا تبدیل موجک[۲]، ویژگی‌های آماری مشخص کننده بافت به صورت اندازه-گیری‌های آماری محلی مانند شش ویژگی بافت تامورا[۳] استفاده می‌شوند. در میان شش ویژگی تامورا- زبری[۴]، جهت[۵]، نظم[۶]، تباین[۷]، شبه خط بودن[۸]، تباین و زبری- سه ویژگی اول بسیار مهم هستند. سه ویژگی دیگر مرتبط به سه ویژگی اول هستند و به کارا بودن توصیف بافت اضافه نمی‌کنند. محدودیت ویژگی‌های تامورا این است که هیچ کاری در چندین وضوح برای به‌دست آوردن مقیاس انجام نمی‌شود.

در کار منژونات و ما[۹] در سال ۱۹۹۷ ویژگی های بافت بر اساس مشخصه بافت پیکسل‌ها یا بلوک‌های کوچک که در ناحیه هستند، استخراج شده است. برای هر ناحیه، ویژگی‌های بافت همه بلوک‌های ۴×۴ که شامل ناحیه می شود، به عنوان ویژگی ناحیه استفاده شده است. مشکل این ویژگی این است که نمی‌توانند به طور موثری مشخصه بافت کل ناحیه را توصیف کنند. یک راه شهودی برای حل این مشکل، توسعه ناحیه شکل- دلخواه[۱۰] به یک ناحیه مستطیلی با اضافه کردن مقادیری خارج از مرز و پس از آن، اعمال تبدیلات بلوک است. اگرچه، چون نواحی در تصاویر دنیای واقعی معمولاً همگی بافت ندارند، این اضافه کردن[۱۱] ابتدایی مولفه‌های ساختگی که ناحیه اصلی را توصیف نکرده و بنابراین کارایی ویژگی بافت به‌دست آمده را کاهش می‌دهند، معرفی می‌شوند. یک راه‌حل دیگر به‌دست آوردن یک مستطیل داخلی[۱۲] (IR) از یک ناحیه تبدیلات بلوکی که می‌توانند برای تولید ضرایب از ویژگی بافت محاسبه شده، انجام شوند. این کار هنگامی که روی بافت ناحیه همگن است و مستطیل داخلی اطلاعات کافی برای توصیف مشخصه بافت ناحیه دربردارد، مناسب است. اگرچه، نواحی تصویر در تصاویر دنیای واقعی معمولاً همگن نیستند. به‌علاوه، در بسیاری موارد، می‌توانیم فقط یک مستطیل داخلی که یک سطح کوچک از ناحیه اصلی را پوشش می‌دهد، به‌دست آوریم. بنابراین ویژگی بافت به‌دست آمده از مستطیل داخلی نمی‌تواند به خوبی مشخصه کل ناحیه را نمایش دهد. برای حل این مسئله، یک الگوریتم کارای استخراج ویژگی بافت برای نواحی شکل- دلخواه توسط لویی و همکاران[۱۳] ارائه شده است[۱۹]. این الگوریتم یک ناحیه شکل- دلخواه را به یک سطح

---

۱ gabor
۲ wavelet transform
۳ Tamura
۴ coarseness
۵ directionality
۶ regularity
۷ contrast
۸ line-likeness
۹ Manjunath & Ma
۱۰ arbitrary-shape
۱۱ padding
۱۲ inner rectangular
۱۳ Lui et al.

مستطیلی با اضافه کردن ابتدایی گسترش می‌دهد. بنابراین یک حلقه نگاشت به مجموعه‌های محدب[1] (Pocs) برای یافتن مجموعه از ضرایب که ناحیه را با نگاشت تکراری بین دامنه تصویر و دامنه تبدیل آن به بهترین حالت توصیف می‌کند، اعمال شده است. در نهایت، ویژگی‌های بافت می‌توانند از ضرایب به‌دست آمده استخراج شوند.

۲-۳-۱-۳- لبه

توصیف هیستوگرام لبه[2] (EHD) برای تاثیر کامل بر بازنمایی تصاویر طبیعی معرفی شده است. این توصیفگر توزیع مکانی لبه‌ها را که در حالت مشابه مانند توصیفگر طرح مکانی است، به‌دست می‌آورد که برای محاسبه توصیفگر هیستوگرام لبه، یک تصویر داده است. ابتدا به زیرتصویرهای ۴×۴ تقسیم شده و هیستوگرام‌های لبه محلی برای هریک از این زیرتصویرها محاسبه شده است. لبه‌ها در حالت کلی به پنج دسته تقسیم می شوند: عمودی، افقی، ۴۵ درجه، ۱۳۵ درجه و خنثی. بنابراین هر هیستوگرام محلی پنج قطعه منطبق با پنج دسته بالا دارد. تصویر به ۱۶ زیرتصویر در نتیجه در ۸۰ قطعه بخش‌بندی شده است. این قطعه‌ها به صورت غیریکنواخت با ۳ بیت/ قطعه در نتیجه در یک توصیفگر به اندازه ۲۴۰ قطعه چندی می‌شوند. اما توصیفگر هیستوگرام لبه نسبت به اعوجاج‌های اشیا یا صحنه بسیار حساس است. هوانگ و دای[3] بردار گرادیان را از تصاویر زیرباند با ترتیب موجک[4] به عنوان ویژگی بافت به‌دست آورده‌اند. بردار گرادیان رویکرد مشابه با توصیفگر گرادیان لبه است.

۲-۳-۱-۴- شکل

شکل، یک مفهوم خوش-تعریف خوب است. ویژگی‌های شکل کاربرد عمومی شامل نسبت نمود[5] ،مدور بودن، توصیفگرهای فوریه، ثابت‌های گشتاور[6] ،قطعه‌های مرزی متوالی و غیره هستند. ویژگی‌های شکل، ویژگی مهمی از تصویر هستند که در کلاس‌بندی تصویر مانند رنگ و ویژگی‌های بافت به طور گسترده استفاده نشده‌اند. ویژگی‌های شکل در بسیاری تصاویر حوزه خاص مانند اشیای مصنوعی مفید هستند. اگرچه، اعمال ویژگی‌های شکل در مقایسه با رنگ و بافت به دلیل عدم دقت در قطعه قطعه‌سازی سخت است، برای تصاویر رنگی بیشترین کاربرد را دارد. علیرغم این دشواری، ویژگی‌های شکل در بعضی سیستم‌ها استفاده شده‌اند و سود بالقوه‌ای را نشان داده‌اند.

---

[1] projection onto convex sets
[2] edge histogram descriptor
[3] Huang & Dai
[4] wavelet
[5] aspect ratio
[6] moment invariants

۲-۳-۲- ویژگی های سطح مفهومی[۱]

ویژگی استخراج شده تصویر بر مبنای رنگ، لبه، شکل و بافت به طور گسترده‌ای در بسیاری کاربردهای بینایی استفاده شده است. اگرچه، «شکاف معنایی» بین ویژگی‌های سطح پایین اولیه و معناهای انسان کارایی کلاس‌بندی را تحت تاثیر قرارداده است. بنابراین شکاف معنایی، پژوهشگران را مجبور به انتقال از سیستم های پیچیده براساس ویژگی سطح پایین به ویژگی های سطح بالای نمایانگرتر با استفاده از منبع مختلفی از مفهوم در تصویر که ویژگی‌های سطح مفهومی نامیده می‌شوند، کردند.

اگرچه اصطلاح «مفهوم» در بینایی محاسباتی، بسیار استفاده شده است، اما تعریف واضحی ندارد. این ویژگی به طور مبهم به عنوان « هر و همه‌ی اطلاعاتی که ممکن است روشی که صحنه و اشیای با آن درک می‌شوند را تحت تاثیر قرار دهد» شناخته می‌شود.

۲-۳-۲-۱- مفهوم محلی

مفهوم محلی رایج‌ترین منبع مفهوم است که مفهوم پیکسل/ وصله[۲]های تصویر اطراف ناحیه موردنظر[۳] را که شامل اطلاعات مفید است، به‌دست می‌آورد. حقه‌ی کلاسیک افزایش اندازه یک آشکارساز بررسی پنجره[۴] که شامل پیکسل‌های اطراف است، یک کاربرد ساده است. همانطور که بیشتر در روش‌های مبتنی بر MRF/CRF به کار رفته‌اند. قطعه قطعه‌سازی تصویر، استخراج مرز شیء، و مدل‌های مختلف شکل/ نقشه[۵] شیء همچنین نمونه‌هایی از مفهوم پیکسل‌های محلی هستند، همانطور که آن‌ها از احاطه‌کننده‌های تصویر مختلف برای تعریف شکل/ مرز آن‌ها استفاده می‌کنند.

استفاده از مفهوم محلی مانند یک بازنمایی سطح معنا در کارهای گسترده اخیر روی کلاس‌بندی صحنه به منظور بررسی شکاف معنایی بین ویژگی‌های سطح پایین و مفاهیم سطح بالا معرفی شده است. عموماً، در سیستم‌هایی که از مفهوم محلی برای کلاس‌بندی صفحه استفاده می‌کنند، مفهوم صحنه با توصیفگرهای محلی همچون کدواژه[۶] به صورت بسته‌کلمات، بسته ویژگی، کدهای بصری توصیف شده است. این روش‌ها معمولاً با بسته‌کلمات که روشی برای تحلیل آمار متن است کار می‌کنند.

---

۱ contextual
۲ patch
۳ region of interest
۴ scanning-window detector
۵ contour
۶ code words

روش بسته‌کلمات[۱] اولین راه پیشنهادی برای تحلیل سند متنی و بیشتر سازگار با کاربردهای بینایی ماشین بود. مدل‌ها با استفاده از یک آنالوگ بصری از یک واژه، به شکل یک بردار ویژگی بصری چندی شده (رنگ، بافت و غیره) به تصاویر، اعمال شده‌اند. فعالیت‌های اخیر نشان داده‌اند که ویژگی‌های محلی با روش بسته‌کلمات برای کلاس‌بندی صحنه که سطوح موثری از کارایی را نشان می‌دهند، بازنمایی شده‌اند.

عموماً ساخت بسته‌کلمات از تصاویر، شامل گام‌های زیر است:

۱) آشکارسازی نواحی/ نقاط موردنظر به طور خودکار
۲) محاسبه توصیفگرهای محلی برای این نواحی/ نقاط
۳) چندی کردن توصیفگرها به واژه‌هایی از لغت‌نامه بصری
۴) پیداکردن رخدادهای هر واژه خاص تصویر در لغت‌نامه برای ساختن بسته‌کلمات (هیستوگرام واژه‌ای).

بعضی از مدل‌های بیزین متنی، مانند تحلیل معنای نهفته احتمالی[۲] (pLSA) و تحلیل دیریلکه نهفته[۳] (LDA) سازگار شده و برای مدل‌کردن دسته‌های صحنه براساس وصله‌های پیکسل‌های محلی استفاده شده‌اند. در تحلیل متن، این مدل‌ها برای کشف عنوان‌ها در یک متن با استفاده از بازنمایی بسته‌کلمات سند استفاده شده‌اند.

در این روش، تصاویر را به عنوان اسناد داریم و عنوان‌ها را به عنوان دسته‌های صحنه جستجو می‌کنیم، بنابراین یک تصویر، مفهوم را به صورت ترکیبی از وصله‌های بصری لبه نشان می‌دهد.

بوش و همکاران[۴] رویکردی ارائه کردند که از دسته‌ای اشیا برای مدل‌کردن صحنه‌های بصری در مجموعه‌های تصویر، براساس ویژگی‌های غیرقابل‌تغییر محلی و pLSA استفاده می‌کند. این روش به طور موفقیت‌آمیزی تا ۱۳ دسته را با دقت ۷۳/۴ درصد کلاس‌بندی می‌کند [۲۰]. کوئل هاس و همکاران[۵] از یک رویکرد مشابه بازنمایی اختلافات به صورت زیر استفاده کردند [۲۱]:

۱) تعداد صحنه‌هایی که کلاس‌بندی می‌شوند (۳ صحنه در کار کوئل هاس و همکاران، توصیفگرهای در سال ۲۰۰۵ و ۱۳ صحنه در کار بوش و همکاران در سال ۲۰۰۶).

---

[۱] bag-of-words
[۲] probabilistic latent semantic analysis
[۳] latent dirichlet analysis
[۴] Bosch et al.
[۵] Quel has et al.

۲) چگونه ویژگی‌ها استفاده شده‌اند: در کار کوئل هاس و همکاران، توصیفگرهای سیفت اطراف یک نقطه موردنظر- توصیفگرهای تنک، در کار بوش و همکاران، توصیفگر سیفت[۱] روی یک شبکه منظم و با استفاده از وصله‌های متحدالمرکز اطراف هر نقطه برای مجاز کردن واریانس مقیاس توصیفگرهای متراکم محاسبه شده‌اند. به‌علاوه، این چنین تشریح شده است که هنگام کار کلاس‌بندی صحنه و به‌خصوص کار با تصاویر طبیعی همچون ساحل[۲] یا زمین‌باز[۳]، توصیفگرهای متراکم، توصیفگرهای تنک را تولید می‌کنند. فی‌فی و پرونا[۴] به طور مستقل دو نوع از LDA را که اولین بار توسط بلی و همکاران[۵] مطرح شد، پیشنهاد کردند، که برای بازنمایی و یادگیری مدل های سندی طراحی شده بود. در این چارچوب، نواحی محلی ابتدا به طرح‌های محلی به اندازه طرح‌های میانی، هم با روش خودکار، هم گذاشتن هر نشانه انسانی یادگیری شده‌اند.

۲-۳-۲-۲- مفهوم معنایی

مفهوم معنایی نوع رویداد[۶]، فعالیت، یا صحنه دیگری از زیر دسته به تصویر کشیده شده را نشان می‌دهد. این مفهوم همچنین ممکن است حضور و مکان (مفهوم مکانی) اشیای دیگر و ارتباط میان مفاهیم یا اشیاء موجود مختلف در یک صحنه داده شده را نشان دهد.

علاوه بر رنگ و بافت، استفاده از موقعیت مکانی همچنین در کلاس‌بندی صحنه و به‌خصوص ناحیه، مفید است. برای مثال «آسمان» و «دریا» می‌توانند رنگ و بافت یکسانی داشته باشند، اما موقعیت‌های مکانی آن‌ها متفاوت است، آسمان معمولاً در بالای تصویر، در حالیکه «دریا» در پایان تصویر است. موقعیت-های مکانی معمولاً به سادگی با واژه‌های بالاتر[۷] ، پایین [۸] و بالا[۹]، طبق موقعیت ناحیه در یک تصویر تعریف می‌شوند. مرکز ثقل ناحیه و مستطیل مرزی کمینه آن، برای به‌دست آوردن اطلاعات موقعیت مکانی استفاده شده‌اند.

---

۱ sift descriptor
۲ coast
۳ open country
۴ Perona
۵ Blei et al.
۶ event
۷ upper
۸ bottom
۹ top

در کار مانژونات و ما، تاون و سینکلیر[۱] و مزاری[۲]، در مرکز مکانی، یک ناحیه برای بازنمایی موقعیت مکانی آن استفاده شده است. به علاوه، در به‌دست آوردن ویژگی‌های معنایی، رابطه مکانی نسبی مهم‌تر از موقعیت مکانی مطلق است. رشته دو بعدی و متغیرهای آن رایج‌ترین ساختار استفاده شده برای بازنمایی روابط جهتی بین اشیا مانند «چپ/ راست» و «پایین/ بالا» است. اگرچه، این روابط جهتی به تنهایی برای بازنمایی مفهوم معنایی تصاویری که روابط توپولوژیکی را انکار می‌کنند، کافی نیستند. رن و همکاران[۳] یک الگوریتم مدل‌کردن مفهوم مکانی که شش رابطه مکانی بین زوج نواحی را در نظر می‌گیرد، مطرح کردند: چپ، راست، بالا، پایین، در تماس[۴] و جلو[۵]. یک روش مطلوب توسط اسمیت و لی[۶] مطرح شده است که از یک قالب ناحیه ترکیبی[۷](CRT) برای تعریف ترتیب نواحی و هر کلاس معنایی که با CRTهای به‌دست آمده از مجموعه‌ای از تصاویر ساده، استفاده می‌کند.

استفاده از وجود یک شیء خاص، روش دیگری برای به کار بردن مفهوم معنایی برای کلاس‌بندی صحنه است، روش‌های مطرح شده اغلب بر اساس ابتدا قطعه قطعه‌بندی تصویر به منظور بازشناسی و مکان‌یابی[۸] شیء هستند. سپس، کلاس‌بندهای محلی برای برچسب‌گذاری نواحی قطعه قطعه‌شده به صورت تعلق به یک شیء (مثلاً آسمان، افراد، اتومبیل، چمن و غیره) استفاده شده‌اند. در پایان، با استفاده از اطلاعات محلی، صحنه سراسری کلاس‌بندی شده است. لو و ساواکیس[۹] یک رویکرد ترکیبی مطرح کردند [۲۲]: ویژگی‌های سطح پایین و معنایی را به صورت یک چهارچوب دانش کلی با هم جمع کردند که یک شبکه بیزین[۱۰] ایجاد کردند. جهت‌های اتصال های بین گره‌ها (متغیرها) در شبکه بیزین رابطه علت و معلول در صحنه را نشان می‌دهند که این اتصالات، احتمالات شرطی استنتاج وجود یک متغیر دیگر را بیان می‌کنند. هرگره می‌تواند ورودی‌ها و خروجی‌های مستقیمی داشته باشد بصورتی که هر تخصیص، رابطه وابستگی آن به گره‌هایی از مبدا ورودی‌ها (والد) و گره‌هایی که به خروجی‌ها (فرزندان) می‌روند، باشد. کارایی این چارچوب با سه کاربرد شامل ادراک معنایی تصاویر مجسم تشریح شده است:

۱) آشکارسازی موضوعات اصلی عکس در یک تصاویر
۲) انتخاب جذاب‌ترین تصویر در یک رویداد

---

[۱] Town & Sinclair
[۲] Mezari
[۳] Ren et al.
[۴] touch
[۵] front
[۶] Smith & Li
[۷] composite region template
[۸] locaization
[۹] Luo & Savakis
[۱۰] bayesian network

۳) کلاس‌بندی تصاویر به صحنه‌های داخلی و خارجی

آخرین کاربرد به طور خاص به مسئله کلاس‌بندی صحنه اشاره می‌کند. کارایی به طور کمی با استفاده از ویژگی‌های سطح پایین ویژگی‌های معنایی به‌هم‌پیوسته (اشیای آسمان و چمنزار) ارزیابی شده است. این مراحل نشان می‌دهند که کارایی کلاس‌بندی می‌تواند به طور قابل توجهی با بکارگیری ویژگی‌های معنایی در فرایند کلاس‌بندی بهبود یابد.

با استفاده از چارچوب بیزین همراه با مفهوم معنایی، اکسوی و کوپرسکی[۱] همچنین تعریف قواعد بصری را معرفی کردند. بازنمایی صحنه با تجزیه تصویر به نواحی نوعی و مدل کردن فعل و انفعالات بین این نواحی به صورت روابط مکانی، به‌دست آمده است. در آغاز، قطعه قطعه‌بندی تصویر با استفاده از یک الگوریتم تقسیم و ادغام[۲] انجام شده است. سپس، روش به صورت خودکار گروه‌های نواحی نمایشگری که صحنه‌های مختلفی را تفکیک و مدل‌های قاعده بصری را می‌سازند، یادگیری می‌کند. بعد از قطعه قطعه‌سازی تصویر به نواحی، ویژگی‌ها استخراج شده و نواحی کلاس‌بندی می‌شوند. در نهایت، بر اساس این کلاس‌بندی محلی، الگوریتم کل تصویر را کلاس‌بندی می‌کند. یک رویکرد مشابه توسط موجسیلویک و روگوتیز[۳] مطرح شده است که ابتدا تصویر را با استفاده از اطلاعات رنگ و بافت برای یافتن شاخص‌های معنایی (پوست، آسمان، آب و غیره) قطعه‌بندی می‌کند. سپس، این اشیا برای شناسایی دسته‌های معنایی (مانند افراد، خارجی، مناظر طبیعی و غیره) استفاده می‌شوند.

علاوه بر این، کلاس دیگری از روش‌های بسته‌کلمات برای همراه کردن مفهوم محلی با مفهوم مکانی برگرفته از ایده‌ای که در تصاویر طبیعی پیچیده، سیستم‌های کلاس‌بندی صحنه می‌توانند بیشتر با استفاده از دانش مفهومی مانند روابط مکانی رایج بین وصله‌های محلی همسایه یا وضعیت دقیق وصله‌ها در صحنه-های مشخص، بهبود یابند. فرگوس و همکاران[۴] دو مدل جدید ABSpLSA و FSI-pLSA را، که pLSA را که شامل به ترتیب وضعیت دقیق و اطلاعات مکانی در یک حالت انتقال و تغییرناپذیری مقیاس است، گسترش دادند. اگرچه این روش برای کلاس‌بندی شی استفاده شده است، می‌تواند به سادگی برای کارهای کلاس‌بندی صحنه هم سازگار شود [۲۳].

---

۱ Aksari & Koperski
۲ split and merge
۳ Mojsilovic
۴ Fergus et al.

۲-۳-۲-۳- مفهوم هندسی سه‌بعدی

مفهوم هندسی سه‌بعدی برای به‌دست آوردن ساختار هندسی سه بعدی یک صحنه یا «طرح مکانی سطح[۱]» استفاده می‌شود که می‌تواند برای استدلال درباره سطوح پشتیبان، تصادم‌ها، نقاط اتصال و غیره استفاده شود. این نوع از مفهوم بیشتر برای ساخت طرح مکانی و استخراج سه بعدی مفید است.

در این فصل با مروری عمیق، مدل‌های پیشنهادی برای کلاس‌بندی صحنه را به دو دسته مختلف تقسیم می‌کنیم: بینایی محاسباتی و رویکردهای شناختی. به طور رسمی‌تر، مسئله کلاس‌بندی صحنه را با بهبود روش‌های استخراج ویژگی و یادگیری ماشین دنبال می‌کنیم. درحالی که، آخرین کارهای کلاس‌بندی صحنه الهام گرفته از شناخت بصری انسان و توانایی مغز در ادراک صحنه‌های دنیای واقعی اطرافش است.

۲-۳-۳- کلاس‌بندی براساس بینایی محاسباتی

با استفاده از رویکردهای بینایی محاسباتی، صحنه‌ها معمولاً می‌توانند در دو مقیاس مختلف استخراج و کلاس‌بندی ویژگی محلی (ابرپیکسل‌ها[۲]، نواحی، اشیاء و بلوک‌ها) و مقیاس سراسری (کلی، به اصطلاح تمام تصویر) پردازش شوند. به علاوه، درمورد ساختار سیستم، برخی سیستم‌ها منبع متفاوتی از دانش به کمک کلاس‌بندها، ویژگی‌ها و فعالیت‌های بینایی مختلف برای بهبود دقت کلاس‌بندی به دست می‌آورند که سیستم چندوجهی[۳] نامیده می‌شود.

۲-۳-۳-۱- کلاس‌بندی مقیاس محلی

در روش‌های مقیاس محلی، کلاس‌بندی صحنه با استفاده از ویژگی‌های استخراج شده از عناصر زیرتصویر مانند ابرپیکسل‌ها، بلوک‌ها، کدواژه‌ها (بسته ویژگی‌ها)، اشیاء، حباب‌ها، بخش‌ها و نواحی انجام می‌شود. نمونه‌ها در محدوده‌ی بافت ساده یا بازنمایی رنگ بلوک‌های کوچک در فضای ویژگی سطح پایین تا تحلیل مفهومی اشیای مختلف در یک صحنه، یادگیری صریح روابط مکانی بین اشیاء و نواحی، یا یک شیء و نواحی همسایه‌اش قرار دارند.

در چارچوب‌های کلاس‌بندی تصویر مطرح شده توسط کارسون و همکاران[۴]، کلاس‌بندی تصویر با بلاب ورلد[۵] که تصویر را به صورت ترکیبی از حباب‌های دقیق در نظر می‌گیرد، انجام می‌شود. اساساً، بعد از قطعه قطعه‌بندی تصویر به نواحی کاملاً مشخص براساس بافت و رنگ، سیستم، تصاویری در مجموعه داده با

---

۱ surface layout
۲ super pixels
۳ multimodal
۴ Carson et al.
۵ blobworld

پیکربندی‌ها و اندازه نواحی سازنده یکسان را جستجو می‌کند. در نتیجه، بازنمایی داخلی که با بلاب ورلد به‌دست آمده معمولاً در سطح متفاوتی از نوحی، اشیاء یا گروهی از اشیاء انجام شده، بنابراین سیستم هنگامی که اشیای مجزا را در یک پس‌زمینه ساده جستجو می‌کند، نسبت به هنگامی که به دنبال دسته‌های خلاصه بیشتری است، بسیار بهتر عمل می‌کند [۲۴].

گرکانی و پیکارد[۱] کلاس‌بندی صحنه را برای دسته‌های شهر و مناظر طبیعی انجام دادند. آن‌ها از یک هرم قابل هدایت چند مقیاسه[۲] برای یافتن جهت‌های اصلی در زیربلوک‌های ۴×۴ تصویر استفاده کردند: اگر زیربلوک‌ها جهت عمودی اصلی قوی با نسبتاً قوی و همچنین جهت افقی دارند، تصویر به عنوان یک صحنه شهر کلاس‌بندی شده است [۲۵].

در مدل پیشنهادی توسط وگل و شیله[۳]، تصاویر به یک شبکه از بلوک‌های محلی ۱۰×۱۰ تقسیم شده است که به یکی از نه کلاس محلی- مفهوم کلاس‌بندی شده است. در گام بعدی، این اطلاعات محلی خلاصه شده و برای دسته‌بندی تصویر استفاده شده است. مفاهیمی که به عنوان تفکیک‌کننده برای صحنه‌های به کار گرفته مشخص شده‌اند، آسمان، آب، چمنزار، تنه‌های درخت، برگ‌ها، مزرعه، صخره‌ها، گل‌ها و شن هستند. همه تصاویر مجموعه داده به صورت دستی با این نه مفهوم برای به‌دست آوردن داده آموزشی و محک، نشانه‌گذاری شده بودند. برای کلاس‌بندی خودکار مفهوم زمانی که رنگ وجود دارد، نواحی تصویر با الحاق از هیستوگرام‌های رنگی HSI ۸۴- قطعه‌ای ، هیستوگرام لبه- جهت ۷۲- قطعه‌ای و ۲۴ ویژگی از ۲ روش براساس ماتریس هم‌رخدادی سطح خاکستری جیست سراسری بازنمایی شده است. اگر رنگ حذف شده باشد، نواحی تصویر با الحاق یک هیستوگرام چگالی ۳۲- قطعه‌ای، یک هیستوگرام لبه- جهت ۷۲- قطعه‌ای و ۲۴ ویژگی از ماتریس هم‌رخدادی سطح خاکستری بازنمایی شده است. با استفاده از این اطلاعات ویژگی سطح پایین، دو کلاسبند ماشین بردار پشتیبان[۴] (SVM) آموزش داده می‌شوند. کارایی کلاس‌بندی روی سطح ناحیه تصویر با حضور رنگ ۷۱/۷ درصد است، درحالی که در کلاس‌بندی نواحی سطح خاکستری، کارایی کلاس‌بندی تا ۶۵/۷ درصد کاهش می‌یابد [۳۸].

همان‌طور که چارچوب کلاس‌بندی دیگری براساس شبکه‌های تقسیم شده محلی منظم است، وگل و شیله رویکردی برای دسته‌بندی صحنه‌های طبیعی دنیای واقعی براساس معنای به‌دست آمده از درجه‌بندی شباهت دسته‌های صحنه ارائه کردند. در این روش، تصاویر با فرکانس رخداد نه مفهوم معنایی محلی،

---

[۱] Gorkani & Picard
[۲] multi scale steerable pyramid
[۳] Vogel & Schiele
[۴] support vector machine

آسمان، آب، چمنزار، تنه‌های درخت، برگ‌ها، مزرعه، صخره‌ها، گل‌ها و شن به عنوان مشخصه‌های دسته صحنه، استخراج شده روی یک شبکه ۱۰×۱۰ منظم دلخواه از نواحی زیرتصویر بازنمایی شده‌اند. برای هر مفهوم معنایی محلی، فرکانس رخداد آن در یک تصویر خاص مشخص شده و هر تصویر با بردار رخداد مفهوم بازنمایی شده است. براساس این مشخصه‌ها، یک بازنمایی نوعی دسته برای هر دسته صحنه یادگیری شده است. بازنمایی دسته ارائه شده مزایای تصمیمات باینری بازنمایی نشده درباره مفاهیم معنایی حاضر یا غایب در تصویر را دارد («بله، صخره‌ها وجود دارند.» درمقابل «خیر، هیچ صخره‌ای وجود ندارد.»). به جای آن، تصمیمات نرمی درباره احتمال این‌که یک مفهوم معنایی خاص وجود دارد، بازنمایی می‌شود. سپس، تصویر معمولاً با محاسبه فاصله ماهالانوبیس[۱] بین بردار رخداد مفهوم تصاویر و بازنمایی نوعی، اندازه‌گیری می‌شود.

فی‌فی و پرونا[۲] رویکردی برای کلاس‌بندی صحنه‌های طبیعی که نیاز به هیچ مهارتی در نشانه‌گذاری مجموعه آموزشی ندارد، ارائه کردند. مجموعه داده به کار رفته در آزمایش‌های آن‌ها شامل سیزده دسته صحنه سطح پایه بود: بزرگراه، دره و شهرها، ساختمان‌های بلند، خیابان‌ها، جنگل‌ها، ساحل، کوه، منظره‌باز، حومه شهر، اتاق خواب، آشپزخانه، اتاق نشیمن و اداره. تصاویر صحنه‌ها به صورت مجموعه‌ای از وصله‌های محلی که به طور خودکار روی نقاط تغییرناپذیر نسبت به مقیاس آشکارسازی شده و با یک بردار ویژگی تغییرناپذیر نسبت به چرخش، روشنایی و زاویه دید سه بعدی توصیف شده، مدل شده بودند. هر وصله با یک کدواژه از یک لغت‌نامه بزرگ از کدواژه‌ها که قبلاً با خوشه‌بندی کا-میانگین[۳] روی مجموعه‌ای از وصله-های آموزشی یادگیری شده، بازنمایی می‌شود. در مرحله یادگیری، مدلی که بهترین توزیع کدواژه‌های موجود در هر دسته صحنه بازنمایی می‌کند، با یک الگوریتم یادگیری بر پایه تخصیص نهفته دیریکله[۴] ساخته شده بود. در مرحله کلاس‌بندی، ابتدا شناسایی همه کدواژه‌ها در تصویر نامشخص انجام شده بود. سپس مدل دسته‌ای که بهترین انطباق را با توزیع کدواژه‌های تصویر آزمایشی دارد، مقایسه احتمال تصویر داده شده هر دسته را استنتاج می‌کند. کارایی به‌دست آمده توسط این الگوریتم به ۶۵/۲ درصد دقت می‌رسد [۲۶].

با استفاده از رابطه مکانی میان واژهای بصری، لازبنیک و همکاران[۵] روش پیشنهادی، یک بازنمایی هرم مکانی تصویر با ساختن روی ایده پیشنهادی که یک هسته انطباق هرمی[۶] برای یافتن یک انطباق تخمینی

---

۱ mahalanobis distance
۲ Fei-Fei & Perona
۳ K-means
۴ latent dirichlet allocation
۵ Lazebnik et al.
۶ pyramid match kernel

بین دو مجموعه از عناصر استفاده شده، استخراج می‌شود. برای نوعی از واژه‌های بصری، ابتدا مشخص می‌کند در چه مکانی واژه بصری در تصویر وجود دارد. سپس در هر سطح از هرم، زیرتصویرهای سطح قبلی به چهار زیرتصویر تقسیم می‌شوند. یک هیستوگرام برای هر زیرتصویر در هرم شامل فرکانس هر قطعه از یک واژه بصری خاص ساخته شده است. در پایان، بازنمایی هرم مکانی تصویر به صورت برداری شامل همه هیستوگرام‌های وزن‌دهی شده رخ داده در سطح انطباق، به‌دست می‌آید. وزن‌های هر هیستوگرام برای جریمه انطباق دو قطعه هیستوگرام منطبق مرتبط به یک زیرتصویر بزرگ‌تر و تایید انطباق زمانی که قطعه‌ها به یک زیرتصویر کوچک‌تر اشاره دارند، استفاده شده است. نویسندگان یک ماشین بردار پشتیبان را با استفاده از قانون یک- دربرابر- همه[۱] برای اجرای بازشناسی دسته صحنه، به کار گرفتند. این روش زمانی که توصیفگرهای سیفت وصله‌های پیکسل ۱۶×۱۶ برای یک شبکه با فضای ۸ پیکسلی که در ساخت لغت‌نامه بصری به کار رفته محاسبه شود، دقت ۸۱/۴ درصد را می‌دهد [۲۷].

در کار لیم و جین[۲] در سال ۲۰۰۴، ماشین‌های بردار پشتیبان روی نواحی از تعداد کمی تصویر متعلق به هفت دسته معنایی آموزش داده شده‌اند. این نواحی که منجر به یک خروجی بالای SVM می‌شوند، خوشه‌بندی شده و یک لغت‌نامه بصری را تشکیل می‌دهند. تصاویر مشاهده نشده با هیستوگرام‌های این لغت‌نامه بازنمایی شده‌اند. این موضوع ارزیابی نشده است که آیا لغت‌نامه بصری در حقیقت شامل خوشه‌های معنایی هست یا خیر. ژانگ و ژانگ[۳] یک لغت‌نامه بصری با آموزش یک نگاشت خود- سازماندهی شده روی بردارهای ویژگی نواحی قطعه قطعه شده تصویر ساخته می‌شود. با مشاهده صفحه دوبعدی به عنوان تصویر دودویی و اجرای عملگر فرسایش[۴]، هر مولفه متصل یک کدواژه بصری را بازنمایی می‌کند [۲۸]. مفهوم معنایی این کدواژه‌ها به کیفیت قطعه قطعه‌بندی ناحیه بستگی دارد. همچنین، فاوکوئر و بوجما[۵] یک لغت-نامه بصری با خوشه‌بندی و گروه‌بندی نواحی قطعه قطعه شده تصویر ساختند. هم، گام قطعه قطعه‌بندی و هم گروه‌بندی براساس رنگ هستند. کاربران می‌توانند نمونه‌های مثبت و منفی از کلید واژه‌های بصری برای بازنمایی «تصاویر ذهنی[۶]» انتخاب کنند. این موضوع آشکار نیست که چقدر انسان‌ها خوب می‌توانند مفهوم معنایی کدواژه‌های بصری را به دلیل اندازه و از دست دادن مفهوم آن‌ها، بازشناسی کنند.

---

۱ one-versus-all
۲ Lim & Jin
۳ Zhang & Zhang
۴ erosion
۵ Fauqueur & Boujemaa
۶ mental images

کومار و هبرت[۱] ساختار مصنوعی را که خانه، حصار و غیره در تصاویر طبیعی است، درک کردند. این رویکرد از یک مشخصه تصادفی چند مقیاسه علّی به عنوان یک مدل ابتدایی روی برچسب‌های کلاس استفاده می‌کند و وابستگی‌های مکانی محلی متفاوت داده ساختاریافته یا نیافته را از طریق یک هیستوگرام چند مقیاسه روی جهت‌های گرادیان مدل می‌کند [۲۹]. در سیستم پیشنهادی توسط تاون و سینکلیر[۲]، شبکه‌های عصبی برای کلاس‌بندی نواحی قطعه قطعه شده تصویر به یکی از یازده کلاس معنایی مانند آجر، ابر، خز یا شن آموزش داده می‌شوند. نواحی تصویر با ویژگی‌های رنگ و بافت بازنمایی و تصاویر با گنجینه‌های بصری بازیابی می‌شوند. اگرچه همه رویکردهای موجود برچسب‌گذاری ناحیه، نیاز به داده آموزشی نشانه‌گذاری شده دارند، تلاش‌ها اغلب باعث نرخ بالای کلاس‌بندی و بنابراین دقت و فراخوانی بالا می‌شوند [۳۰].

بونت و ویلا[۳] یک رویکرد متفاوت برای بازنمایی صحنه مطرح کردند که تصویر با یک بردار ویژگی با ابعاد بالا به‌دست آمده از خروجی درختی از فیلترهای غیرخطی بازنمایی شده است. سیستم آن‌ها براساس شباهت بین نواحی با ویژگی‌های بافتی، مکانی یا رنگی خاص (مانند اتومبیل‌های مسابقه‌ای، غروب آفتاب‌ها) کلاس-بندی می‌کند. اما روش آن‌ها، براین اساس است که یک امضاء[۴] با ابعاد خیلی بالا وجود دارد، شکل‌گیری یک بازنمایی معنادار داخلی صحنه را مجاز نمی‌کند.

لی و همکاران[۵] در سال ۲۰۰۴ یک نسخه شبه- نظارتی[۶] جدید از الگوریتم EM[۷] را برای یادگیری توزیع‌های کلاس‌های شیء توسعه دادند. تصاویر به صورت مجموعه‌ای از بردار ویژگی چندین نوع از نواحی خلاصه بازنمایی شده‌اند. هر ناحیه خلاصه به صورت ترکیبی از توزیع‌های گاوسی روی فضای ویژگی‌اش مدل شده است. نواحی به کار رفته در کلاس‌بندی می‌توانند از پردازش‌های قطعه‌بندی مختلفی حاصل شوند که به عنوان «ناحیه خلاصه[۸]» شناخته می‌شوند. یک بخش کلیدی این رویکرد این است که نیاز به مشخص بودن موقعیت اشیای هر تصویر نیست. آزمایش‌ها روی مجموعه‌ای از ۸۶۰ تصویر کارایی رویکرد را نشان می‌دهد [۳۱].

لی و همکاران در سال ۲۰۰۵، یک رویکرد یادگیری دو فازی تولید/ تفکیک کننده مطرح کردند که می‌توانست اشیاء را با انواع ویژگی چندتایی بازشناسی کند. هدف این کار توسعه یک روش کلاس‌بندی برای

---

۱ Kumar & Hebert
۲ Town & Sinclair
۳ Bonet & Voila
۴ signature
۵ Li et al.
۶ semi-supervised
۷ expected maximization
۸ abstract region

کلاس‌بندی خودکار تصاویر صحنه خارجی است. عبارت تولیدکننده طول توصیف تصاویر را نرمال می‌کند، که می‌تواند تعداد دلخواهی از ویژگی‌های استخراج شده هر نوع داشته باشد. در مرحله تفکیک‌کنندگی، یک کلاس‌بند یاد می‌گیرد تصاویر همان‌طور که با این توصیف طول- ثابت بازنمایی شده‌اند، شامل شیء هدف هستند. نتایج عملی آن‌ها با استفاده از رنگ، بافت و ویژگی‌های ساختار، کارایی بازیابی محتمل روی ۳۱ دسته شیء ابتدایی و ۲۰ مفهوم سطح بالا را نشان می‌دهند [۳۲].

رشته دو بعدی و متغیرهای آن رایج‌ترین ساختار استفاده شده برای بازنمایی روابط جهتی بین اشیاء مانند «چپ/ راست»، «پایین/ بالا» هستند. اگرچه، این روابط جهتی به تنهایی برای بازنمایی مفهوم معنایی تصاویر صرف‌نظر از روابط توپولوژیکی کافی نیستند. برای پشتیبانی بهتر بازیابی تصویر مبتنی بر معنا، یک روش خوب توسط اسمیت و لی ارائه شد. این سیستم از یک قالب ناحیه ترکیبی (CRF) برای تعریف ترتیب مکانی نواحی و هر کلاس معنایی که توسط این قالب از مجموعه تصاویر ساده مشخص شده، استفاده می‌کند.

لیپسون و همکاران[۱] یک رویکرد مدل‌سازی مفهوم مکانی در فضای شیء ارائه کردند که مدل کردن صحنه مبتنی‌بر پیکربندی[۲] برای شاخص‌گذاری مبتنی‌بر مفهوم و کاربردهای بازیابی نامیده می‌شود. آن‌ها روابط فتومتریک و کیفی بین اشیای مختلف صحنه را در یک حالت مکانی مدل کردند و این روابط را برای استخراج دیگر صحنه‌ها با مفهوم معنایی مشابه استفاده کردند. مزیت اصلی سیستم آن‌ها این است که صحنه با مفهوم رنگی مشابه همانند تصویر پرس‌وجو با سیستم آن‌ها انتخاب نشده است، درحالی که، صحنه‌های از نظر معنایی مشابه با اشیای با رنگ متفاوت می‌تواند بازیابی شوند. مدل‌های صحنه به شدت نسبت به طرح مکانی صحنه خاص هستند (اقیانوس بالای شن متفاوت با اقیانوس کنار شن است) و مدل‌های چندین صحنه ممکن است برای هر نوع صحنه نیاز باشد. به علاوه، مدل صحنه براساس تصویر پرس‌وجو ساخته شده و به دلیل این‌که تنها، دیگر تصاویر مشابه معنایی خواسته شده‌اند، این مدل می‌تواند برای تصویر پرس‌وجو ساده و خاص باشد.

مسئله دنبال شده توسط بوش و همکاران برای کشف اشیاء در هر تصویر با یک حالت غیرنظارتی و استفاده از توزیع اشیاء برای اجرای کلاس‌بندی صحنه بود. برای این هدف، تحلیل معنایی نهفته احتمالی به بازنمایی بسته‌کلمات بصری هر تصویر اعمال شده بود. یک لغت‌نامه بصری برای بسته کلمات بصری استخراج توصیفگر سیفت را روی یک دامنه رنگی HSV ارائه شده مدل می‌کند. طبق معمول، کا- میانگین برای ساختن لغت‌نامه به کار گرفته شده است. کلاس‌بندی صحنه روی توزیع شیء با یک کلاس‌بند کا- ان

---

[۱] Lipson et al.
[۲] configuration-based

ان[1] انجام شد. ترکیب (غیرنظارتی) pLSA دنبال شده با (نظارتی) کلاس‌بندی نزدیک‌ترین همسایه مطرح شده در کار بوش و همکاران در سال ۲۰۰۶ روش‌های قبلی را اجرا می‌کنند. به عنوان مثال، دقت این رویکرد ۸/۲ درصد بهتر از روش پیشنهادی در کار فی‌فی و پرونا در سال ۲۰۰۵ است که روی مجموعه داده یکسانی انجام شده است.

اغلب رویکردها برای یادگیری دسته‌های بصری شیء نیاز به صدها تصویر آموزشی دارند. در کار فی‌فی و همکاران در سال ۲۰۰۴، یک الگوریتم بیزین افزایشی برای یادگیری مدل‌های تولیدکننده دسته‌های شیء از تنها تعداد کمی تصاویر آموزشی توسعه دادند. این روش از اطلاعات قبلی به‌دست آمده از دسته‌های شیء که قبلاً یادگیری شده استفاده می‌کند. یک مدل احتمالی تولیدکننده برای بازنمایی شکل و ظاهر مجموعه‌ای از ویژگی‌های متعلق به شیء استفاده شده است. پارامترهای مدل به صورت افزایشی با یک مدل بیزین یادگیری شده‌اند. الگوریتم روی تصاویر ۱۰۱ دسته شیء کاملاً متفاوت شامل چهره، کامپیوتر شخصی، توت فرنگی، گورخر، فنجان، صندلی و غیره.

۲-۳-۳-۲- کلاس‌بندی مقیاس سراسری

انسان‌ها می‌توانند صحنه‌های بصری پیچیده‌ای را در یک نگاه، بدون توجه به تعداد اشیاء با وضعیت‌ها، رنگ‌ها، سایه‌ها بافت‌های متفاوت که ممکن است در صحنه باشند، بازشناسی کنند. بنابراین، برخی مطالعات عملی پیشنهاد شدند که بازشناسی صحنه‌های دنیای واقعی ممکن بود با استفاده از پیکربندی سراسری، صرف‌نظر از جزئیات درباره مفاهیم محلی و اطلاعات شیء آغاز شوند. با استفاده از تنها ویژگی سطح پایین در مقیاس سراسری و صرف‌نظر از دانش مفهومی، با دقت پایین در مسئله کلاس‌بندی صحنه نتیجه می‌کند. بنابراین چارچوب‌های دیگر مبتنی‌بر موارد سراسری انتقال یافته با استفاده از ویژگی‌های سطح پایین غیرمفید به ویژگی‌های سراسری به‌دست آمده از ادراک انسان هستند. در این بخش، تعدادی از کارهای پیشنهادی را که از ویژگی سطح پایین در مقیاس سراسری برای کلاس‌بندی صحنه استفاده می‌کنند، مرور می‌کنیم.

رنینگر و مالیک[2] یک بازنمایی سراسری از صحنه برای بازشناسی دسته آن به کار گرفتند. پایه استفاده از بازنمایی سراسری این بود که ویژگی‌های سراسری روی زمینه بصری انسان پردازش کرده‌اند و نیازی به تحلیل ویژگی‌های سراسری، که به انسان اجازه بازشناسی سریع دسته صحنه را می‌دهد، نیست [۳۳]. برای بررسی این‌که انسان‌ها می‌توانند به سرعت و موازی با زمینه بصری، بافت را پردازش کنند، از یک بازنمایی

---

[1] K-NN
[2] Renninger & Malik

سراسری براساس تکستون[۱] استفاده شده است. فرآیند اصلی استفاده شده برای کدگذاری بافت‌ها با ساخت یک لغت‌نامه از الگوهای مجزا شروع شده، می‌تواند ویژگی‌ها و ساختارهای بافت‌های متفاوت موجود در صحنه‌ها را شناسایی کند. لغت‌نامه با خوشه‌بندی کا- میانگین روی مجموعه‌ای از پاسخ‌های فیلتر ساخته شده است. با لغت‌نامه ساخته شده هر تصویر به صورت هیستوگرام فرکانس تکستون‌ها بازنمایی شده است. تصاویر صحنه به کار رفته در آزمایش‌ها، ده دسته سطح پایه بودند: ساحل، کوه، جنگل، شهر، مزرعه، خیابان، حمام، اتاق خواب، آشپزخانه و اتاق نشیمن. کارایی مدل پیشنهادی حدوداً ۷۶ درصد باقی می‌ماند.

تانگ و چنگ[۲] از یک الگوریتم یادگیری فعال ماشین بردار پشتیبان برای استنتاج بازخورد ارتباط موثر برای کلاس‌بندی تصویر استفاده کردند. الگوریتم، نزدیک‌ترین تصاویر به پرسش کاربر را انتخاب و به سرعت یک مرز را که تصاویری را که مفهوم پرس‌وجوی کاربر را از باقیمانده مجموعه داده ارضا می‌کند، یادگیری می‌کند. برای یادگیری مفهوم تصویر، این روش از هردو ویژگی‌های مبتنی‌بر رنگ و بافت در یک مدل چند وضوحی استفاده می‌کند. ایده استفاده از هردو ویژگی رنگ و بافت در درجه متفاوت وضوح این است که نوع متفاوتی از تصاویر صحنه می‌توانند با ویژگی‌های بصری مختلف در گام‌های وضوح متفاوت مشخص شوند. هنگام بازنمایی یک تصویر با ویژگی رنگ، در بالاترین وضوح، آن‌ها رنگ را با استفاده از یک ماسک رنگی ۱۲ بیتی مشخص می‌کنند. برای ذخیره اطلاعات رنگی وضوح‌های پایین‌تر، هشت ویژگی اضافی هیستوگرام‌های رنگی و پراکندگی‌ها در کانال‌های H، S و V، کشیدگی و توسعه رنگ، از کل تصویر به صورت سراسری محاسبه شده است. علاوه بر ویژگی رنگ، تبدیل موجک گسسته[۳] (DCT) در سه سطح مقیاس برای یافتن میانگین انرژی، پراکندگی، کشیدگی و گستردگی بافت تصویر استفاده شده است. درنهایت، یک بردار ۱۴۴ بعدی تولید شده که برای کلاس‌بندی تصویر استفاده شده است. نتایج عملی نشان می‌دهند که چارچوب پیشنهادی تا دقت ۹۴ درصد بعد از پنج بار آموزش SVM_{active} بالا می‌برد [۳۴].

وایلایا و همکاران[۴] از بازنمایی سطح پایین سراسری تصویر براساس رنگ و لبه‌ها به خصوص برای کلاس‌بندی شهر (بناهای تاریخی، پل‌ها، خیابان‌ها و ساختمان‌ها) درمقابل فضای باز (جنگل‌ها، زمین‌های کشاورزی، سواحل و کوه‌ها) استفاده کردند. برای انتخاب یک مجموعه ویژگی تفکیک‌کننده برای کلاس‌بندی تصویر شهر درمقابل فضای باز، تعدادی از ویژگی‌های برجسته تصویر براساس رنگ (هیستوگرام رنگ، بردارهای انسجام رنگ)، بافت (گشتاورهای ضرایب DCT بافت) و لبه (هیستوگرام جهت لبه و بردارهای انسجام جهت لبه) در این کار آزمایش شده‌اند. قدرت تفکیک هر ویژگی براساس مقادیر فاصله درون-

---

۱ texton
۲ Tong & Chang
۳ discerete wavelet transformation
۴ Vailaya et al.

کلاسی[۱] و بین- کلاسی[۲] ارزیابی شده است. براساس کلاس‌بند کا- ان ان روی هریک از پنج ویژگی سطح پایین، نتایج عملی کلاس‌بندی شهر- فضای باز نشان می‌دهد که بهترین دقت ۹۳/۹ درصد با یک کلاس‌بند ۵- ان ان با استفاده از بردارهای انسجام جهت لبه به‌دست آمده است [۳۵].

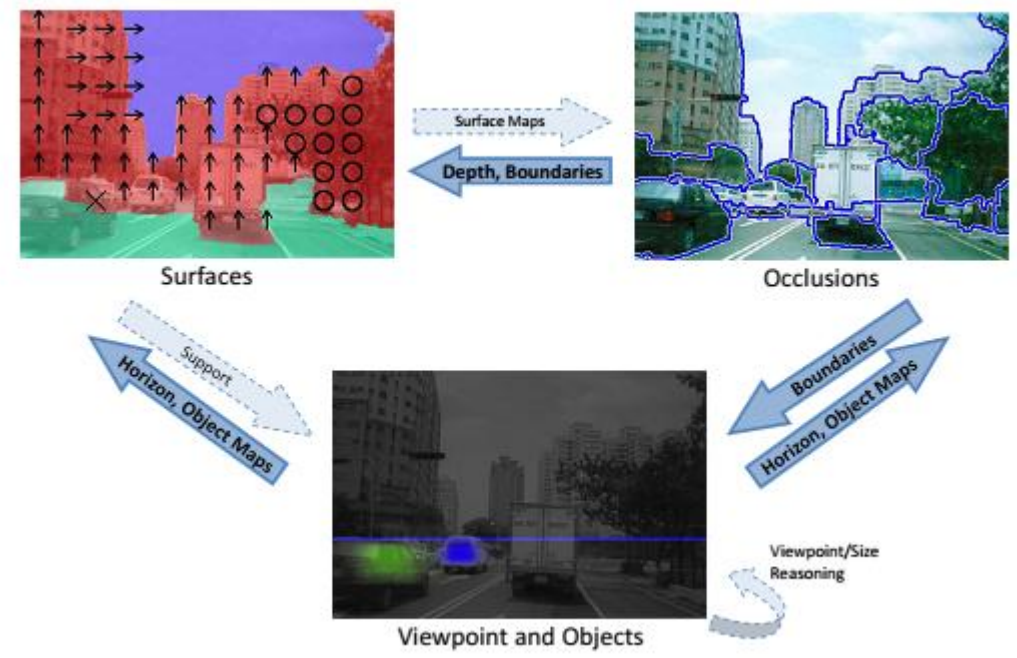
۲-۳-۳-۳- کلاس‌بندی صحنه چند وجهی

یک سیستم کلاس‌بندی صحنه چند وجهی شواهد ارائه شده با چندین منبع اطلاعاتی را جمع می‌کند. در یک سیستم چند وجهی، اطلاعات می‌توانند در سطح متفاوتی از ترکیب جمع شوند: سطح ویژگی‌ها، سطح وظایف[۳] و سطح کلاس‌بندها. ترکیب در سطح ویژگی‌ها می‌تواند با دو یا بیشتر ویژگی سازگار به شکل یک بردار ساده یا پیکربندی یک چارچوب یادگیری مانند MRF سلسله مراتبی یا شبکه بیزین روی مجموعه متفاوتی از ویژگی‌ها همچون بافت و رنگ در ویژگی سطح پایین یا شی‌ای با طرح مکانی سه بعدی در سطح مفهومی انجام شود. اجتماع در سطح وظایف، الهام گرفته از این حقیقت است که یک سیستم ادراک صحنه کامل نیاز به هماهنگی فعالیت‌های زیادی همچون استدلال انسداد[۴]، تخمین جهت سطح، بازشناسی شیء، قطعه قطعه‌سازی و دسته‌بندی تصویر دارد. هریک از این وظایف بعضی جنبه‌های صحنه را بررسی می‌کند، اما این فعالیت‌ها مرتبط هستند، آن‌ها مشخصه‌هایی از صحنه یکسانی را بازنمایی می‌کنند. بنابراین، یک فعالیت می‌تواند مشخصه‌های (خروجی‌ها) معناداری برای فرآیند یادگیری فعالیت دیگری ایجاد کند. در سطح کلاس‌بندها، ترکیب با اجتماع مجموعه‌ای از کلاس‌بندهایی که با یک یا بیشتر نوع متفاوت از ویژگی آموزش داده شده انجام شده است. کلاس‌بندها ممکن است در پیکربندی متوالی یا موازی در کنار یک کلاس‌بند پایانی متصل شوند. درمیان این سه سناریو اجتماع کامل در چارچوب‌های کلاس‌بندی چند وجهی، ترکیب در سطح وظایف اخیراً توجه زیادی جلب کرده است. در این بخش، برخی از سیستم‌های چند وجهی مهم‌تر را که براساس این سه سطح از اجتماع مطرح شده‌اند، مرور می‌کنیم.

هوئیم و همکاران[۵] نگران این بودند که چگونه زیرفعالیت‌های مرتبط با ادراک صحنه می‌توانند در روشی که کارایی هریک از آن‌ها براساس تصاویر طبیعی بهبود می‌یابد، جمع شوند. هر تصویر طبیعی یک نگاشت ثبت شده است که یک مشخصه از صحنه را توصیف می‌کند. اجتماع زیرفعالیت به تخمین جهت‌های صفحه ادراک صحنه سه بعدی، مرزهای انسداد، اشیاء، زاویه دید دوربین و عمق نسبی و عمق نسبی اعمال شده است. مدل کلی این چارچوب در شکل ۲-۱ نشان داده شده است [۳۶].

---

۱ intra-class
۲ inter-class
۳ tasks
۴ occlusion
۵ Hoeim et al.



شکل ۲-۱ تعامل مفهومی میان زیروظایف [۳۶].

۲-۴- کلاس‌بندی صحنه براساس شناخت بصری

دسته‌ی دیگری از پژوهش‌ها تلاش کردند کشف کنند که چگونه انسان‌ها صحنه‌های اطراف را درک و شناخت بصری را برای کلاس‌بندی صحنه مدل می‌کنند. چندین نقطه شروع نگرانی بینایی انسان دیدگاهی به این‌که کدام دسته‌های صحنه برای انسان مرتبط هستند؟ چگونه انسان آن‌ها را توصیف می‌کند یا کدام ویژگی‌های صحنه برجسته هستند. کدام ویژگی‌های تصویر قابل ارزیابی توسط انسان هستند؟ چگونه اشیای ضعیف و قوی می‌توانند بر دقت درک صحنه تاثیر گذارند؟

روش‌های پیشنهادی الهام گرفته از درک انسان برای کلاس‌بندی صحنه تلاش می‌کنند یک مدل مفهومی از کل تصویر یادگیری کنند که یک بازنمایی کلی از صحنه را که به عنوان جیست شناخته شد تولید می‌کند. به‌طور دقیق‌تر، ویژگی‌های تصویر به نواحی یا اشیاء گروه‌بندی نمی‌شوند، بلکه در یک چارچوب کلی صحنه-محور نشان داده می‌شوند. فعالیت‌های اخیر نشان داده‌اند که توصیفگرهای معنایی برای صحنه‌های دنیای واقعی می‌توانند با این بازنمایی‌های کلی بدون نیاز به قطعه‌بندی تصویر به دست آیند. این مشاهده با حجم زیاد شواهد به دست آمده از مطالعات روانشناسی و علوم شناختی سازگار است.

روگوویتز و همکاران[1] مجموعه کاملی از آزمایش‌ها به منظور تعیین دسته‌های معنایی تصاویری با توصیفگرهای شفاهی نمایشی برای این دسته‌ها انجام دادند. نتایج اولیه آن‌ها روی مجموعه داده ۹۷ تصویری که انسان‌ها تصاویر را روی دو محور مصنوعی در مقابل طبیعی و انسان در مقابل غیرانسان دسته‌بندی کرده بودند، بود. در یک آزمایش متوالی با ۱۹۶ تصویر، آن‌ها ۴ دسته اصلی و ۲۰ زیردسته شناسایی کردند. اگرچه، برای هر دسته معنایی، از مجموعه‌ای از توصیفگرهایی که انسان‌ها برای توصیف این دسته‌ها پیدا کردند، استفاده کردند. برای مثال،«آب»، «آسمان/ ابرها»، «برف» و «کوه‌ها» به عنوان موارد بسیار مهم برای دسته‌های طبیعی بودند.

همچنین، به نظر می‌رسید انسان‌ها نسبت به حضور اشخاص در تصویر بسیار حساس هستند. ترکیب رنگی و ویژگی‌های رنگی هنگام مقایسه تصاویر طبیعی مهم هستند، اما به ندرت در توصیف تصاویر شامل اشخاص یا محیط‌های ساخته دست انسان استفاده شده است. خطوط مستقیم، مرزهای مستقیم و لبه‌های تیز، مشخصه‌های تصاویر مصنوعی هستند درحالی‌که تصاویر صحنه‌های طبیعی مرزهای سخت و توزیع یکنواختی از لبه‌ها دارند.

اولیوا و تورالبا صحنه‌ها را به عنوان یک موجودیت واحد شناختند که می‌توانند با ویژگی‌های ادراکی برآورد شده از مجموعه‌ای از آماره‌های مرتبه دوم سطح پایین بازنمایی شوند. بازنمایی ادراکی، پوشش مکانی صحنه، براساس یک مطالعه تجربی است که از انسان‌ها درباره ۸۱ خوشه تصویر طبق شباهت صحنه و توصیف انتخاب آن‌ها در کلمات پرسیده شده است. گام اول تقسیم ۸۱ تصویر به دو گروه بود. در گام دوم، موضوع‌ها، هر دو گروه را به دو زیرگروه دیگر تقسیم کرده و در گام سوم، موضوع‌ها هریک از چهار گروه را به دو زیرگروه تقسیم کرده و درمجموع ۸ زیرگروه حاصل شد. در پایان هر گام، موضوع‌ها برای تشریح معیاری که آن‌ها در چند واژه به کار بردند، پرسیده می‌شد. در نتیجه، مجموعه‌ای از ویژگی‌های ادراکی همچون طبیعی بودن، باز بودن، سخت بودن، وسعت و ناهمواری است.

ایده اصلی این آزمایش این است که صحنه‌های با ویژگی‌های پوشش مکانی مشابه، دسته معنایی یکسانی را به اشتراک گذاشته و بنابراین اطلاعات معنایی سطح بالا می‌توانند مستقیماً براساس ویژگی‌های سطح پایین برای بازنمایی یک صحنه همانند یک شیء واحد با یک شکل منحصر به‌فرد تخمین زده شوند. شکل ۲-۲ نمونه‌ای از برچسب‌گذاری مجموعه‌ای از تصاویر آزمایشی براساس مقایسه پوشش مکانی آن‌ها با تصاویر آموزشی است.

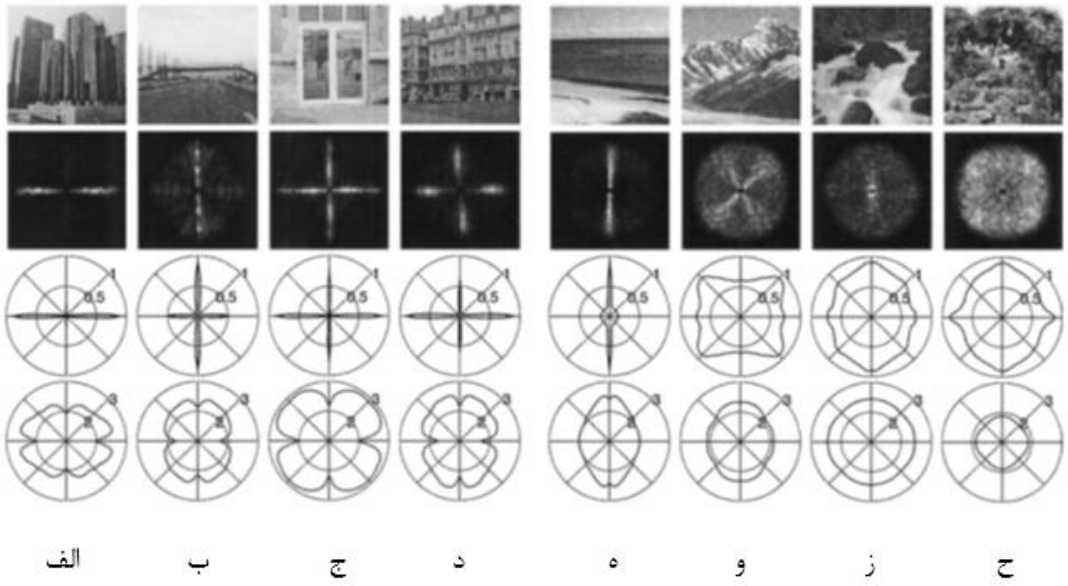
---

[1] Rogowitz et al.

شکل ۲-۲ الف- دید مسطح از محیط مصنوعی، ساختار عمودی، فضای کوچک با عناصر بزرگ، ب- دید مسطح محیط از محیط شهری مصنوعی نیمه- بسته، ج- دید مصطح از محیط شهری مصنوعی بسته، فضای بزرگ با عناصر کوچک، د- دید دورنما از محیط شهری بسته مصنوعی، فضای بزرگ با عناصر کوچک، ه- دید مسطح از محیط شهری مصنوعی، ساختار عمودی [۲].

به جای بازنمایی یک تصویر در دامنه مکانی، اولیوا و تورالبا پوشش مکانی صحنه را براساس تبدیل فوریه گسسته تصویر به صورت تابعی فاز تبدیل فوریه و طیف اندازه تصویر تعریف کردند. تابع فاز، اطلاعات مربوط به ویژگی‌های محلی تصویر، شکل و وضعیت مولفه‌های تصویر را نشان می‌دهد. درمقابل، طیف اندازه اطلاعات محلی نشده درباره ساختار تصویر می‌دهد: جهت، هموار بودن[۱]، طول و عرض کناره‌هایی که تصویر صحنه را می‌سازند. مربع اندازه تبدیل فوریه (طیف انرژی) توزیع انرژی سیگنال درمیان فرکانس‌های مکانی مختلف را می‌دهد. بنابراین، طیف انرژی، یک بازنمایی تغییرناپذیر نسبت به ترتیب و شناسه‌های اشیاء از صحنه است که فقط الگوهای ساختاری اصلی در تصویر را کدگذاری می‌کند. برای بازنمایی روابط مکانی بین ساختارهای اصلی در تصویر، تورالبا و اولیوا از توزیع مکانی اطلاعات طیفی توصیف شده با تبدیل فوریه پنجره‌ای شده[۲] استفاده می‌کنند. بنابراین، یک شکل کلی از صحنه با استفاده از هردوی طیف انرژی و طیف نگاره[۳] بازنمایی شده است. شکل ۲-۳ امضاهای طیفی هشت دسته صحنه را نشان می‌دهد.

۱ smoothness
۲ windowed fourier transform
۳ spectrogram



شکل ۲-۳ نمونه‌های دسته صحنه‌های مختلف، طیف انرژی و امضاهای طیفی آن‌ها به ترتیب.
تصاویر از الف به ح این صحنه‌ها را نشان می‌دهند: ساختمان بلند، بزرگراه، دید نزدیک شهری، بزرگراه، دید نزدیک شهری، مرکز شهر، ساحل، کوه، دید نزدیک طبیعی و جنگل‌ها [۲].

با الهام از ایده پوشش مکانی صحنه، اولیوا و تورالبا یک بازنمایی صحنه- محور برای ایجاد یک توصیف معنایی از تصاویر دنیای واقعی در چندین سطح دسته‌بندی ارائه کردند. بازنمایی صحنه- محور براساس ترکیب ویژگی‌های پوشش مکانی توصیف‌کننده شکل صحنه (مانند محدوده عمق، باز بودن، وسعت، انحراف از خط افق، عمود بودن) و ماهیت مفهوم آن (مانند طبیعی بودن، مشغول بودن[۱] و سخت بودن). برای بازنمایی فضای صحنه، یک بازنمایی جزئی از تصویر با استفاده از فیلترهای گابور در مقیاس‌ها و جهت مختلف استفاده شده است.

$$A_x^2(x,k) = \{|i(x) * g_k(x)|^2 \downarrow M \qquad ۱-۲$$

$i(x)\downarrow$ تصویر ورودی و $g_k(x)$ پاسخ ضربه یک فیلتر گابور است. شاخص k فیلترهای تنظیم شده با مقیاس‌ها و جهت‌های مختلف را شاخص‌گذاری می‌کند. نشانه $\downarrow M$ عملگر زیرنمونه‌برداری در دامنه مکانی بازنمایی می‌کند که بازنمایی نتیجه $A_M(x, k)$ یک وضوح مکانی از M^2 پیکسل دارد. بنابراین، یک مدل احتمالی برای هر دسته برای یادگیری روابط بین بازنمایی جزئی هر دسته و توصیفگر فضا (محدوده عمق، بازبودن، طبیعی بودن، مشغول بودن و غیره) مطرح شده است. به طور مشابه، اولیوا و تورالبا در سال ۲۰۰۶

[۱] busyness

یک طرح برای یادگیری مجموعه‌ای از ویژگی‌های سراسری (طبیعی بودن، باز بودن، وسعت، عمق، سختی، پیچیدگی، انحراف نسبت به خط افق، تقارن) طبق به پیکربندی آماری جهت و فرکانس‌های مکانی ارائه شده است. هر مقدار ویژگی سراسری، یک ترکیب وزن‌دار از اندازه خروجی بانکی از فیلترهای جهت‌دار چند مقیاسه است که می‌تواند با مجموعه‌ای از قالب‌های ویژگی‌های سراسری بازنمایی شوند. قالب‌های سراسری اندازه‌گیری می‌کنند که هر تصویر صحنه چه اندازه طبیعی، باز، وسعت داده شده، سخت و غیره است.

پژوهش در ادراک صحنه همچنین درگیر این پرسش است که انسان به کجای تصویر نگاه می‌کند و کدام ویژگی تصویر را ما احتمالاً استفاده می‌کنیم. هندرسون و فریرا[۱] یک مرور کامل بر حقایق پایه درباره شناخت بصری دید صحنه و حرکات چشم و اشیاء ضعیف درمقابل قوی در تصاویر انجام دادند.

به‌علاوه، بوسول[۲] در یک مطالعه اولیه نشان دادند که بینندگان بر نواحی خالی، یکنواخت و غیرمفید صحنه که نواحی «موردنظر» صحنه اغلب منطبق با سطوحی با مفهوم فرکانسی مکانی بالا و چگالی لبه، تباین محلی بالا هستند، تمرکز نمی‌کنند. این مطالعه مطرح کرد که انسان‌ها به سرعت روی هر شیء مصنوعی در یک صحنه تمرکز می‌کنند چون اشیاء مصنوعی معمولاً لبه‌ها و گوشه‌های تیز و بنابراین فرکانس‌های مکانی بالا دارند.

پژوهش بیشتر روی این پرسش که کدام ویژگی‌های تصویر عموماً توسط انسان برای اجرای بازشناسی صحنه استفاده شده، انجام شده است. مک‌کاتر و همکاران[۳]، آزمایش‌های بازشناسی را روی هشت دسته صحنه (بزرگراه، خیابان، درون شهر، ساختمان بلند، ساحل، منظره باز، جنگل و کوه) اجرا کردند. با تحلیل طیف فاز دسته‌های صحنه، آن‌ها یافتند که نواحی تشخیصی دسته- خاص در طیف فاز باعث دسته‌بندی‌های درست می‌شود. نویسنده‌ها پیشنهاد کردند که این جهت‌ها و پهناهای باند تشخیصی شامل اطلاعات صحنه با بیشینه‌کردن دقت درون کلاسی و کمینه کردن بین کلاسی هم‌پوشانی هستند.

۲-۴-۱- مولفه‌های پایه درک صحنه

بازنگری اخیر از یک ساختار محاسباتی برای پردازش بصری در اولیه توسط رنسیک[۴] انجام شده است و به عنوان یک نقطه شروع برای تحلیل فعلی استفاده می‌شود. در ساختار مثلثی ذهن رنسیک ویژگی‌های بصری سطح پایین موازی با کل فیلد بصری تا سطح پیچیدگی نوع- شیء نامیده می‌شود. واسط میان ویژگی‌های ساده همچون لبه‌ها و گوشه‌ها و بازنمایی‌های شیء پیچیده یک شاخه از پردازش متوالی با

---

۱ Henderson & Ferreira
۲ Buswell
۳ McCatter et al.
۴ Rensik

محاسبه‌ای که تنظیمات[۱] نامیده می‌شود که یک تحلیل معنایی خام عادلانه از ماهیت صحنه (جیست آن) را شامل می‌شود. مثلاً آیا این یک خیابان شلوغ، یک آشپزخانه یا یک ساحل و طرح کلی مکانی سخت آن است. در شاخه‌ای دیگر، توجه به بخش فضایی کوچک ورودی بصری را انتخاب کرده و به طور گذرا اراده نوع-شیء را به بازنمایی‌های سازگار اشیای مورد توجه محصور می‌کند. سپس این اشیای را با جزئیات بیشتر پردازش شده، تشخیص هویت ویژگی‌شان صورت می‌گیرد.

توجه بصری اغلب با راهنمای بصری که ذهن ما از طریق آن دنیا را می‌بیند و توجه کلاس‌بندی شده به چندین نوع براساس این‌که آیا آن‌ها شامل حرکت چشم هستند یا نه (آشکار[۲] در مقابل پنهان[۳]) را انتقال می‌دهد و در درجه اول با ویژگی‌های صحنه یا اراده هدایت شده است.

اولین ساختار محاسباتی بیولوژیکی قابل قبول صریح برای کنترل توجه بالا به پایین[۴] توسط کخ[۵] و اولمان[۶] در سال ۱۹۸۵ مطرح شد. در مدل آن‌ها، چندین نگاشت ویژگی (مانند رنگ، شدت نور) در موازات به زمینه بصری محاسبه شده و با یک نگاشت برجسته ترکیب شده است. پس با یک فرایند انتخاب به ترتیب توجه را به سمت مکان‌هایی در جهت کاهش برجستگی آن‌ها می‌برد. در اینجا ما یک ساختار مشابه برای شاخه توجه پردازش بصری فرض کرده و شرح می‌دهیم چگونه این شاخه با مدل کردن اثر کار روی توجه می‌تواند بهبود یابد. انتخاب ویژگی‌هایی که ممکن است توجه پایین به بالا هدایت کند بصورت گسترده در زمینه‌های پژوهشی بصری مطالعه شده است.

در مراحل اولیه پردازش بصری، وظیفه، فعالیت عصبی را با افزایش پاسخ‌های نورون‌های تنظیم شده با مکان و ویژگی‌های یک محرک تعدیل می‌کند. علاوه بر این، آزمایش‌های روانشناسی نشان داده است که شناخت این هدف به تقویت برجستگی آن کمک می‌کند. به عنوان مثال، خطوط سفید عمودی برجسته‌تر می‌شوند، اگر ما به دنبال آن‌ها باشیم. مطالعه اخیر حتی نشان می‌دهد که دانش بهتر از هدف منجر به جستجوی سریع‌تر می‌شود. به عنوان مثال، دیدن یک تصویر دقیق ازهدف بهتر از دیدن یک تصویر هم‌نوع معنایی و یا همان گروه به عنوان هدف است. این مطالعات اثرات بایاس ویژگی‌های هدف را نشان می‌دهد. (به عنوان مثال، تریسمن و گلید[۷] نشان داده اند که جستجو برای حروف ربط ویژگی‌ها (به عنوان مثال، جستجوی ارتباط رنگ x جهت: یافتن مورد قرمز عمودی در میان موردهای قرمز افقی و سبز عمودی)

---

۱ setting
۲ overt
۳ covert
۴ bottom-up
۵ Koch
۶ Ullman
۷ Treisman & Gelade

کندتر از «pop-out» می‌باشد (به عنوان مثال، پیدا کردن یک آیتم به رنگ سبز در میان وسایل قرمز رنگ). این مشاهدات محدودیت‌ها را روی روش‌های بایاس ممکن اعمال کرده و امکان ایجاد ویژگی‌های ترکیبی جدید را در پرواز (به عنوان ترکیبی از ویژگی‌های ساده) از بین می‌برد. یک مدل محبوب به دست آوردن بایاس ویژگی بالا به پایین[۱] و رفتار جستجوی بصری جستجوی هدایت شونده[۲] است. این مدل دارای معماری پایه یکسان است که توسط کخ و اولمان پیشنهاد شد. اما، علاوه بر این، بایاس مبتنی بر ویژگی با وزن‌دهی نگاشت‌های ویژگی به شیوه‌ای از بالا به پایین به‌دست می‌آید. به عنوان مثال، در وظیفه آشکارسازی یک نوار قرمز، نگاشت ویژگی حساس به رنگ قرمز وزن بیشتری دارد، از این رو، ساخت نوار قرمز برجسته‌تر است. یکی از سوالاتی که باقی می‌ماند این است که چگونه بطور بهینه وزن ویژگی‌های نسبی مانند بیشینه کردن قدرت شناسایی مجموعه‌ای از اهداف رفتاری موردنظر در میان پارازیت تنظیم می‌شود.

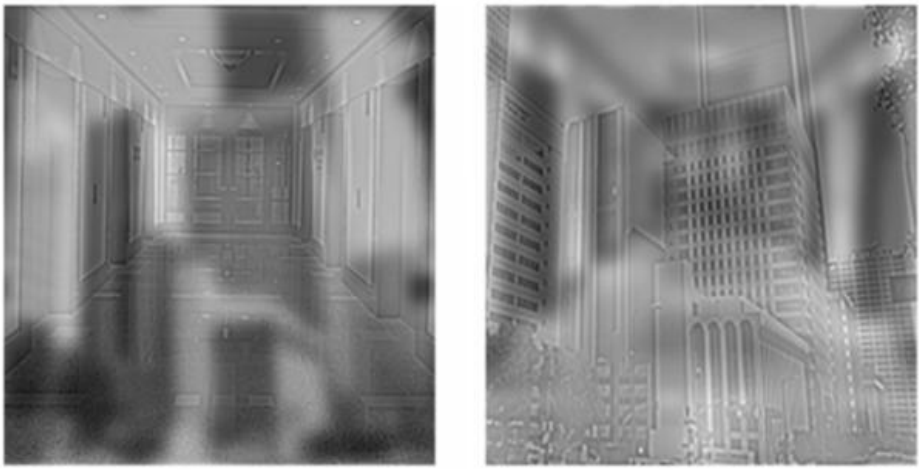
با توجه به یک نگاشت برجسته، مدل‌های مختلفی ارائه شده است که محل حضور بعدی، از جمله اشکال مختلف را به عنوان برنده همه‌ی (بیشینه-انتخابگر[۳]) الگوریتم‌ها را می‌برد. پس با داشتن کانون توجه انتخابی، تشخیص موجودیت در مکان صحنه مهم است. بسیاری از مدل‌های تشخیص که پیشنهاد شده‌اند، می‌توانند براساس عواملی از جمله انتخاب شکل‌های هندسی اولیه (به عنوان مثال، جت‌های گابور، شکل‌های هندسی اولیه مانند جئون‌ها، تکه‌های تصویر یا حباب‌ها، و واحدهای تنظیم نمایش)، فرآیند تطبیق (به عنوان مثال، انطباق پیوند پویای خود سازماندهی، انطباق احتمالی)، و عواملی دیگر کلاس‌بندی شوند.

تشخیص با مسئله حفظ اطلاعات بصری دنبال می‌شود. یک نظریه محبوب، نظریه فایل هدف حافظه ترانس- حرکت[۴]، فرض می‌کند هنگامی که توجه به سمت یک شی می‌رود، ویژگی‌های بصری و اطلاعات محلی به یک فایل شی در حافظه کوتاه مدت بصری که در سراسر حرکت‌هایش نگهداری شده، محدود می‌شود. آزمایش‌های روانشناسی بیشتر نشان داده شده که تا سه یا چهار فایل شیء ممکن است در حافظه نگهداری شود. مطالعات پژوهشی در مورد لایه‌های عصبی از حافظه مشغول به کار درمورد پستانداران و انسان‌ها نشان می‌دهد که قشر پیشانی و شیاردار ممکن است هر دو عملکردی و آناتومی به یک حافظه‌ی «چه چیزی» برای ذخیره‌سازی ویژگی‌های بصری از محرک، و یک حافظه‌ی «کجا» برای ذخیره‌سازی اطلاعات مکانی تجزیه می‌شود. برای به خاطر سپردن محل اشیاء، ما در اینجا فرضیه‌های قبلی یک نگاشت برجسته را به منظور پیشنهاد یک توپوگرافی دوبعدی نگاشت وظیفه-ارتباط[۵] (TRM) است که وظیفه-ارتباط، موجودیت‌های صحنه را کدگذاری می‌کند، گسترش می‌دهیم. در شکل ۲-۴ به ترتیب در قسمت‌های

---

[۱] top-down
[۲] guided search
[۳] maximum-selector
[۴] object file theory of trans-saccadic memory
[۵] task-relevance

الف و ب یک بازنمایی ترکیبی از یک صحنه ورودی با فرکانس مکانی بالا و یک صحنه شهر با فرکانس زمانی پایین و مکمل صحنه ترکیبی آورده شده است.



شکل ۲-۴ الف- یک بازنمایی ترکیبی از یک صحنه ورودی با فرکانس مکانی بالا و یک صحنه شهر با فرکانس زمانی پایین. ب- مکمل صحنه ترکیبی [۱۰].

اخیراً تورالبا بازنمایی کلی از صحنه براساس ویژگی‌های پوششی مکانی[۱] و غیره که تحلیل اشیای مولفه را انشعاب داده صحنه را به‌عنوان یک شناسه واحد نمایش می‌دهد. این رویکرد جیست را به‌عنوان یک بردار از ویژگی‌های مفهومی فرموله می‌کنند. با پردازش به صحنه حاشیه نویسی شده، این نویسنده رابطه بین مفهوم صحنه و دسته‌ی اشیایی که در آن صحنه رخ می‌دهد و از جمله ویژگی‌های شی همچون مکان، اندازه یا مقیاس را درک کرده و از آن برای تمرکز کانون توجه به مکان‌های هدف محتمل استفاده می‌شود. این مسئله یک نقطه شروع خوب برای مدل کردن نقش جیست در توجه راهنما ایجاد می‌کند. چون جیست به سرعت محاسبه شده است، می‌تواند به‌عنوان راهنمای آغازی توجه ذخیره شود. اما، به صورت متوالی، TRM پیشنهادی ما که مداوماً به‌روزرسانی شده ممکن است به عنوان یک راهنمای بهترذخیره شود. به عنوان مثال، درصحنه‌های پویا مانند صحنه‌های ترافیک که محیط مداوماً درحال تغییر است و هدف‌هایی همچون اتومبیل‌ها و عابران پیاده اطراف آن درحال حرکتند، جیست ممکن است بدون تغییر باقی بماند و بنابراین ممکن است خیلی مفید نباشد مگر به عنوان یک راهنمای آغازی.

[۱] naturalness

استفاده از توجه راهنما را مکان‌های هدف محتمل، رویکردهای مبتنی‌بردانش را برای مدل کردن حرکات چشم در مقابل رویکردهای مبتنی برتصویر[۱] ترغیب می‌کند. یک رویکرد مشهور مشابه، نظریه بررسی[۲] است که پیشنهاد می‌کند این توجه اغلب با یک روش بالابه‌پایین مبتنی‌برمدل داخلی صحنه هدایت شده است. مدل‌های بینایی ماشینی یک رویکرد مشابه برای تشخیص اشیاء به کارگرفته‌اند. به عنوان مثال، ری‌بک وهمکاران[۳] اشیاء را با استفاده از باز اجرای دنباله‌ای از حرکات چشم و انطباق ویژگی‌های تصویر به‌طور واضح تشخیص می‌دهند.

به‌طورخلاصه، ما مولفه‌های ساختاری پایه‌ای را ترغیب کرده‌ایم‌که معتقدیم برای درک صحنه ضروری هستند. به عنوان مثال، یکی از بهترین نمونه‌های سیستم‌های تحلیل صحنه بلادرنگ تبدیل‌کننده‌ی بصری[۴] (VTRA)، یک سیستم بینایی ماشینی که تفسیرهای‌کلامی بلادرنگ را در هنگام تماشای یک بازی فوتبال تلویزیونی تولید می‌کند. سیستم بصری سطح پایین‌آن‌ها تشخیص می‌دهد و همه‌ی اشیای قابل دیدن از یک بالاسری (چشم‌پرنده) زاویه‌ی دید را دنبال‌کرده و یک بازنمایی هندسی از صحنه ادراک شده تولید می‌کند (۲۲بازیکن، زمین و مکان‌های‌گل). پس این بازنمایی میانی با دنباله‌ای ازشبکه‌های باور بیزین[۵] که روابط مکانی را ارزیابی، رویدادهای حرکت موردنظر برنامه‌ها ومنظورها را به صورت افزایشی تشخیص می-دهد، تحلیل شده است. مدل، شامل یک خلاصه، نماد غیربصری برجستگی که هر رخداد تشخیصی براساس تازگی، فرکانس، پیچیدگی، اهمیت بازی و دیگر عوامل را مشخص کرده است. درنهایت، سیستم یک تفسیرکلامی تولید می‌کند که معمولا با آغاز هر رویدادی که تشخیص داده شده اما ممکن است در میان باشد، شروع می‌شود، اگرچه رویدادهای برجسته قبل ازتکمیل جمله‌ی فعلی، رخ می‌دهند. هنگامی که این سیستم رویدادهای بسیار موثری را در زمینه‌ی کاربرد خاصی دریافت می‌کند، به دلیل پیچیدگی محاسباتی آن به یک محیط ساخته شده سطح بالا و یک وظیفه‌ی خاص محدود شده و نمی‌تواند به یک مدل ادراک صحنه عمومی گسترش یابد. درواقع، انسان‌های مشابه که به‌طور انتخابی اشیاء مرتبط درصحنه را درک می‌کنند، VTRA رسیدگی کرده و مداوماً روی همه‌ی اشیاء نظارت کرده و تلاش می‌کند همه‌ی عوامل معلوم را به‌طور همزمان تشخیص دهد. رویکرد مطرح شده در اینجا با VTRA متفاوت است نه فقط در

---

[۱] image-based
[۲] scanpath
[۳] Rybak et al.
[۴] visual- translation
[۵] bayesian belief networks

این‌که هیچ چیز در مدل ما وجود ندارد که آن را برای یک محیط یا وظیفه‌ی خاص قراردهد. به علاوه، ما فقط اشیاء و رویدادهایی را که انتظارداریم مرتبط با وظیفه درحال انجام باشد، نگهداری می‌کنیم.

۲-۴-۲- دسته‌بندی سریع پایه- سطح صحنه

ناظران انسانی می‌توانند معنای یک تصویر جدید را اگر فقط یک تثبیت واحد داده شده باشد، درک کنند. در طول این نگاه گذرا، یک مجموعه غنی از اطلاعات، کمیت‌های سطح همچون رنگ و بافت اشیاء و طرح کلی مکانی تا ویژگی‌های عملیاتی و مفهومی فضای صحنه و حجم استنباط می‌کنیم. در واقع، از یک توصیف کوتاه مفهومی صحنه همچون «جشن تولد»، ناظران می‌توانند حضور یک انطباق تصویر را که توصیف هنگامی که در یک جریان بازنمایی بصری ترتیبی سریع (RSVP) اعمال شده و ۱۰۰ میلی‌ثانیه مشاهده شده، آشکار کنند. همچنین این توصیف کوتاه به عنوان دسته پایه- سطح برای صحنه بصری در نظر گرفته شده است و به رایج‌ترین برچسب استفاده شده برای توصیف یک مکان را اشاره می‌کند.

دسته‌بندی اصلی الانور راش[۱] وهمکاران نشان داده که ناظران انسانی ترجیح می‌دهند از پایه- سطح برای توصیف اشیاء استفاده کرده و زمان‌های عکس‌العمل کوتاهتری را برای نامگذاری اشیاء در پایه- سطح بیشتری نسبت به فرمانروایی یا فرمانفرمایی نشان می‌دهند. فرض شده است که پایه- سطح دسته‌بندی برتری داده شود زیرا تشابه با دسته و پراکندگی بین- دسته‌ای را بیشینه می‌کند.

در حوزه‌ی صحنه‌های بصری، اعضای دسته‌ی پایه- سطح مشابه، تمایل به داشتن ساختارهای مکانی مشابه و تهیه اعمال محرک مشابه دارد. برای مثال، اغلب محیط‌های معمولی هم‌چون «جنگل‌ها» مکان‌های ضمیمه‌شده‌ای را که با درختان و شاخ و برگ احاطه شده است، نمایش می‌دهند. یک تصویر مکان مشابه از نزدیک ممکن است «پوست درخت» یا «خزه» و از دور «کوه» یا «حومه‌ی شهر» نامیده شود. بنابراین، مشخصه‌ی طرح کلی مکانی یک صحنه، اعمالی را که می‌تواند در صحنه رخ دهد، محدود می‌کند. یک «جنگل» موجب مجموعه‌ی محدودی از پیاده‌روی شود که یک «حومه‌ی شهر» ممکن است گزینه‌های بیشتری برای جهت یابی را موجب شود زیرا فضا باز است. علیرغم این‌که این ویژگی‌های عملیاتی و ساختاری جزء ماهیت معنای صحنه هستند، نقش آن‌ها در شناخت صحنه هنوز مشخص نشده است.

۲-۴-۳- رویکرد شیء- محور برای تشخیص بصری سطح بالا

بسیاری از مدل‌های موثر تشخیص بصری سطح بالا شی- محور، اشیای رفتاری و بخش‌ها، کوچکترین اجزای تحلیل صحنه هستند. در این دیدگاه، معنای یک صحنه واقعی از موجودیت‌های مجموعه‌ای از اشیاء

---

[۱] Eleanor Rosch

شامل آن، در طی آزمون هم رخدادی شی و ترتیب‌دهی مکانی یادگیری شده، دنیای ظاهر می‌شود. به طور متناوب، شناسایی یک یا چند شی مهم ممکن است برای فعالسازی طرح صحنه کافی باشد و بنابراین، تشخیص را آسان می‌کند. اگر چه رویکرد شی- محورکلید رویکردهای رسمی و محاسباتی درک صحنه در ۳۰ سال گذشته بوده است، پژوهش در شناخت بصری چالش‌هایی را برای این دیدگاه مطرح کرده است، به‌خصوص هنگامی که مراحل پردازش بصری و توانایی ما در تشخیص صحنه‌های جدید در یک نگاه گذرا را شرح می‌دهد. تحت شرایط دید بهبود یافته، همچون وضوح مکانی کم یا هنگامی که فقط مرزهای تنک نگهداری می‌شوند، ناظران انسانی هنوز می‌توانند یک دسته‌ی پایه- سطح صحنه را تشخیص دهند. با این محرک‌ها، اطلاعات شناخت شیء که نمی‌تواند به‌صورت محلی پوشانده شود، کاهش می‌یابد. این نتایج پیشنهاد می‌کند که اطلاعات شناخت صحنه ممکن است قبل از یک تحلیل جزئی اشیاء کامل شود، به‌دست آید. بنابراین، پژوهش با استفاده از تغییر الگوهای بی‌بصیرت نشان می‌دهد که ناظران حساس به آشکارسازی تغییرات اشیاء و نواحی محلی در یک صحنه تحت شرایطی که معنای صحنه ثابت می‌ماند، هستند.

در رویکردهای شیء- محور، اشیاء و بخش‌ها[1] به عنوان اتم‌ها (عناصر پایه) در نظر گرفته می شوند که صحنه‌ها را می‌سازند. در این دیدگاه، معنای صحنه‌های دنیای واقعی از شناس‌ه‌های مجموعه‌ای از اشیایی که در آن وجود دارند، تجربه هم‌رخدادی شیء و ترتیب مکانی یادگیری شده‌اند، به‌دست می آید. همچنین، شناسایی یک یا چند شیء برجسته می‌تواند برای فعال کردن طرحی از صحنه کافی باشد و بنابراین بازشناسی را ساده کند.

برخلاف این‌که این رویکرد کلید رویکردهای رسمی و محاسباتی ادراک صحنه در سی سال گذشته است؛ پژوهش در شناخت بصری چالش‌هایی به این دیدگاه وارد کرده است، به‌خصوص هنگام تشریح مراحل اولیه پردازش بصری و توانایی انسان در بازشناسی صحنه‌های جدید با یک نگاه گذرا در شرایط دید ضعیف مانند وضوح مکانی پایین، یا هنگامی که فقط طرح‌های تنک[2] نگه داشته می‌شوند. این نتایج پیشنهاد می‌کنند که اطلاعات شناسه صحنه ممکن است قبل این‌که یک از تحلیل جزئی‌تر اشیا کامل شود، به‌دست آید. بنابراین، پژوهش با استفاده از تغییر الگوهای کور که نشان می‌دهند ناظران نسبت به آشکارسازی تغییرات اشیاء و نواحی محلی در یک صحنه تحت شرایطی که معنای صحنه ثابت باقی می‌ماند، نسبتاً حساس هستند. در پایان، هنوز مشخص نشده است آیا اشیایی که در یک نمای مختصر صحنه درک شده‌اند،

---

[1] part
[2] sparse

می‌توانند شناسایی شوند یا از طریق ادراک دیگر اطلاعات بصری هم‌رخداد مانند ویژگی‌های سطح پایین، ثابت‌های توپولوژیکی یا اطلاعات بافت استنتاج شوند.

درنهایت، هنوز مشخص نشده است که آیا اشیایی که می‌توانند در یک صحنه حاضرکه به طور مختصر درک شده یا با درک اطلاعات بصری هم‌رخدادی دیگر همچون ویژگی‌های سطح پایین، ثوابت بیولوژیکی یا اطلاعات بافت استنباط شده، شناسایی شوند یا خیر.

۲-۴-۴- رویکرد صحنه- محور برای تشخیص بصری سطح بالا

یک مجموع متناوب از تحلیل رویکرد صحنه- محورکه صحنه کامل به‌عنوان کوچکترین جزء تشخیص سطح- بالا عمل می‌کند. با این چارچوب، بازنمایی بصری آغازی با سیستم بصری در سطحی که کل صحنه و نه اشیای قطعه قطعه شده، با هرصحنه به عنوان یک شکل واحد رفتار می‌کند، ساخته شده است. به جای اجزای اولیه‌ی بصری مبتنی‌بر بخش و هندسه‌ی محلی، این چارچوب فرض می‌کند که ویژگی‌های سراسری بازتابی از ساختار صحنه، طرح کلی و عملی که می‌تواند به‌عنوان ابتدایی‌ها در دسته‌بندی صحنه عمل می‌کند. یک فعالیت رسمی نشان داد که صحنه‌ها، عضویت دسته پایه- سطح یکسان را که طرح کلی مکانی مشابهی دارند، به اشتراک می‌گذارند. به‌عنوان مثال، راهرو، یک فضای باریک طولانی با دسترسی چشم انداز بالاست، همان‌طور که جنگل مکانی با یک بافت متراکم سراسری است. مدل‌کردن اخیر موفقیت در شناخت صحنه‌های پیچیده دنیای واقعی درسطوح پایه و بالا از ویژگی‌های سطح پایین به‌طورنسبی (مانند جهت، بافت و رنگ) یا ویژگی‌های طرح مکانی پیچیده‌تر مانند بافت، عمق میانگین و چشم‌انداز بدون نیاز به شناسایی اولیه اشیای مولفه، نشان داده است.

اگرچه این‌که کدامیک از ناظران انسانی از این ویژگی‌های سراسری در تشخیص صحنه‌ها استفاده می-کنند، هنوز نامعلوم است. یک رویکرد صحنه- محور شامل پردازش سراسری و کلی است. اگر این پردازش یک بازنمایی حساس به طرح کلی و ساختار یک صحنه بصری را تولید کند، پردازش سراسری است.

اثرتقدم سراسری موثر نشان داد که ناظران به شکل سراسری محرک حذف سلسله مراتبی حساس‌تر از حروف مولفه آن‌هاست. به‌طور قابل توجهی، اثر تقدم سراسری به خصوص برای محرک شامل الگوهای با عنصر زیاد، همانند حالتی در بیشترین صحنه‌های دنیای واقعی، قوی است. یک توالی از پردازش سراسری، توانایی استخراج آماره‌های ساده با سرعت و دقت یا ذخیره‌ی اطلاعات از نمایش‌گر است. به عنوان مثال، اندازه میانگین عناصردر یک مجموعه‌ی دقیق و خودکار درک شده است. همانند جهت میانگین عناصر

جانبی و بعضی از توصیفگرهای بافت تباین[۱]، همانند مرکز جرم گروهی ازاشیاء، بازنمایی‌های سراسری ممکن است به‌طورتضمینی یادگیری شوند. همان‌طور که ناظران می‌توانند به‌طورتضمینی از طرح‌های سراسری آموزشی برای ساده‌سازی جستجوی بصری استفاده کنند. همان‌طورکه همه‌ی این نتایج، اهمیت ساختار و روابط سراسری را نشان می‌دهد، یک تعریف عملیاتی از تحلیل سراسری صحنه‌های دنیای واقعی که از بین رفته است، بسیاری از ویژگی‌های دنیای واقعی می‌توانند در طبیعت سراسری و کلی باشند. به‌عنوان مثال، تعیین سطح پارازیت یک اتاق یا درک تقارن کلی فضا، تصمیمات کلی هستند که نمی‌توانند از تحلیل محلی به تنهایی به‌دست آیند، اما، به تحلیل نسبی چندین ناحیه نیاز دارد.

در مقابل، رویکردهای صحنه- محور، اطلاعات سراسری صحنه را مطرح می‌کنند، که رسماً «الگو[۲]» یا «قاب‌ها[۳]» و اخیراً «جیست» نامیده شده‌اند و براساس ترتیب مکانی کلی اشیای حاضر با مفهوم پس‌زمینه به‌دست آمده از ویژگی‌های سطح پایین صحنه، استخراج شده‌اند. در این دیدگاه، کل صحنه به عنوان عنصر بازشناسی سطح بالا در نظر گرفته شده است. با این چارچوب، بازنمایی بصری آغازی با سیستم بصری که در سطح کل صحنه و بدون قطعه قطعه‌سازی اشیاء است و با هر صحنه طوری رفتار می‌کند به‌طوری‌که صحنه شکل منحصر به‌فردی داشته باشد. به جای هندسه محلی و اولیه‌های بصری مبتنی‌بر بخش، این چارچوب فرض می‌کند که ویژگی‌های سراسری، ساختار صحنه، طرح مکانی و عملکردی را که می‌تواند به عنوان اولیه‌هایی برای دسته‌بندی صحنه عمل می‌کند، انعکاس می‌دهند. کارهای رسمی نشان داده‌اند که صحنه‌هایی که عضویت دسته سطح پایه یکسانی به اشتراک می‌گذارند تمایل به داشتن طرح مکانی یکسانی دارند. برای مثال، یک راهرو باریک است؛ فضای محدود با یک چشم‌انداز زیاد در حالی که یک جنگل مکانی با دقت کاملاً متراکم است. برخی از فعالیت‌های مدل کردن، موفقیت در شناسایی صحنه‌های پیچیده دنیای واقعی هم در اندازه زیاد و هم‌سطح- پایه‌ها از ویژگی‌های سطح پایین نسبی (مانند جهت، بافت و رنگ) یا ویژگی‌های طرح مکانی پیچیده‌تر همچون بافت، عمق میانگین و چشم‌انداز، بدون نیاز به ابتدا شناسایی اشیای مولفه را نشان داده‌اند. اگرچه، توسعه این‌که ناظران انسانی از این ویژگی‌های سراسری در بازشناسی صحنه استفاده می‌کنند، هنوز نامعلوم است. مقایسه دیدگاه شیء- محور در مقابل صحنه- محور در شکل ۲-۵ نشان داده شده است. شبهه میان رویکردهای صحنه- محور و شیء- محور در ادراک صحنه با این حقیقت که جیست یک صحنه می‌تواند به سرعت مشخص شود، حتی زمانی که تصویر صحنه مات شده

---

۱ contrast
۲ schemas
۳ frames

است، حقیقت این‌که فاصله سریع‌تر از این‌که نیاز به شناسایی همه اشیای سازنده و روابط مکانی آن‌ها به طور متوالی، پرسش آشکار این است که دقیقاً چگونه درک سریع صحنه انجام می‌شود، آشکار شده است.

ابتدا، باید یک شیء تشخیصی[1] به سرعت شناسایی شود (یا چندین شی به سرعت به موازات هم شناسایی شوند) و این‌که جیست صحنه از این شی یا چندین شی و روابط مکانی آن‌ها استناج شوند. بنابراین، حضور یک بخاری، یا یک بخاری و یک یخچال، احتمال وجود یک آشپزخانه را نشان می‌دهد. یک مسئله بالقوه برای این نوع مدل این است که جیست می‌تواند از یک تصویر کاهش‌یافته مشخص شود که هیچ یک از اشیای مولفه نمی‌توانند به سادگی مستقیماً شناسایی شوند. با وجود آن، شاید حتی در این حال، حمایت متقابل و اشیای تشخیص کاهش‌یافته برای صحنه دیگری، فعال‌سازی جمعی کافی برای غلبه بر تجزیه را بالا ببرد.

الف ب

شکل ۲-۵ الف- شیء- محور: آسمان، ساختمان، اشخاص، اتومبیل‌ها، درختان، جاده‌ها؛ صحنه- محور: فضای بزرگ، مصنوعی، صحنه نیمه بسته؛ ب- شیء- محور: لامپ، مبل، پنجره، میز، صحنه- محور: فضای کوچک، مصنوعی و صحنه بسته [۲].

فرض دیگر این است که اطلاعات مکانی فرکانس پایین ممکن است برای تعیین جیست صحنه به کار روند. در دنباله‌ای از مطالعات، اسچینز[2] و اولیوا فلیترهای فرکانسی مکانی بالا و پایین- گذر را به مجموعه‌ای از تصاویر صحنه اعمال کردند. تصاویر فیلتر شده حاصل از زوج صحنه‌ها برای تولید تصاویر در مولفه‌های فرکانسی مکانی پایین یک صحنه که با مولفه‌های فرکانسی مکانی بالای دیگری متصل شده‌اند، ترکیب

[1] diagnostic
[2] Schgns

شده‌اند. این تصاویر ترکیبی به عنوان صحنه‌ها، با توجه به این‌که فرکانس‌های مکانی بیننده به آن تمایل دارند، درک می‌شوند. اسچینز و اولیوا این پرسش مطرح کردند که آیا بینندگان می‌توانند برای استفاده از یک محدوده فرکانسی مکانی دیگر برای تفسیر یک صحنه در مراحل آغازی مشاهده، تحت تاثیر قرار گیرند. مولفه‌های فرکانس مکانی، سیستم بینایی را که معمولاً در ابتدا برای تعیین جیست استفاده می‌کنند، انجام می‌دهند. با اولین تجربه، بینندگان برای استفاده از مولفه فرکانسی مکانی پایین در طول مراحل اولیه شناسایی صحنه، تحت تاثیر قرارگرفته‌اند. بنابراین، شرکت‌کنندگان در شناسایی صحنه‌ها با استفاده از اطلاعات مکانی فرکانس پایین، زمان مشاهده میلی‌ثانیه کاملاً دقیق بودند. پاسخ کار، پایان بازی داشت چون بینندگان می‌توانستند شناسه صحنه‌ها را از یک مجموعه نامتناهی به صورت بالقوه تولید کنند. در دومین مجموعه از آزمایش‌ها، اولیوا و اسچینز نشان دادند که استفاده از فرکانس‌های مکانی پایین در شناسایی سریع صحنه الزامی نبود؛ هنگامی که شرکت‌کنندگان تعلیم داده شدند که از مولفه‌های مکانی فرکانس بالای تصویر ترکیبی استفاده کنند، تمایل به شناسایی صحنه با آن مولفه داشتند. بنابراین، به نظر می رسد که بینندگان طبیعتاً برای استفاده از اطلاعات مکانی فرکانس پایین در گام‌های اولیه شناسایی صحنه تحت تاثیر قرار گرفته‌اند، اما خیلی به آن نیازی ندارند.

همچنین این مدرک وجود دارد که رنگ در شناسایی سریع صحنه، حداقل زمانی که رنگ برای دسته صحنه، تشخیصی است، نقش دارد. از بینندگان درخواست شد که دسته هر صحنه را نام‌گذاری یا تایید کنند. در مراحل نام‌گذاری و تایید، صحنه‌ها برای ۱۲۰ میلی ثانیه به صورت یا رنگی معمولی یا رنگی غیر معمولی یا سطح خاکستری نمایش داده شدند و شرکت‌کنندگان، صحنه‌ها را نام‌گذاری کردند. برای بررسی تشخیصی بودن رنگ برای دسته، حضور رنگ معمولی کار را آسان می‌کرد و حضور رنگ غیر معمولی موجب بازشناسی بر پایه سطح خاکستری شد. برای صحنه‌هایی که رنگ، ویژگی تشخیصی دسته نبود، هیچ تاثیری از رنگ مشاهده نشد. رنگ معمولی، هنگامی که رنگ، ویژگی تشخیصی دسته بود؛ باعث بهبود دسته‌بندی صحنه‌های فیلتر شده پایین‌گذر شد. در نتیجه، یک سازماندهی دقیق از حباب‌های رنگی تشخیصی که می‌توانند برای دسته‌بندی صحنه کافی باشند، پیشنهاد شد.

محاسبات شیء یا صحنه- محور شبیه عملیات تکمیلی هستند که توانایی درک شده از شناسایی صحنه در پایان نگاه گذرا را بالا می‌برد (۲۰۰- ۳۰۰ میلی‌ثانیه). به‌طور آشکار اشیاء، اغلب موجودیت‌هایی هستند که روی صحنه عمل می‌کنند، موجودیت‌های آن‌ها مرکز درک صحنه هستند. اگرچه، بعضی مطالعات نشان داده است که پردازش اطلاعات شیء محلی ممکن است به نمایش تصویری بیشتری که برای شناسایی دسته صحنه مورد نیاز است، نیازمند باشد. دراین مطالعه، ما توسعه‌ی این‌که چه رویکرد صحنه- محور سراسری می‌تواند تفسیر شود و مرحله اولیه کاربرد دسته‌بندی صحنه سریع انسانی را پیش بینی کند، را

آزمایش می‌کنیم. فراتر از اصل تشخیص «جنگل قبل از درختان[1]» این کار دنبال می‌شود تا مفهوم «سراسری بودن» را برای دسته‌بندی سریع صحنه عملی کرده و مقداری جدید از این‌که چگونه ناظران انسانی می‌توانند مکان را به‌عنوان یک «جنگل» شناسایی کنند بدون تشخیص اولیه‌ی «درختان» فراهم می‌کنند.

۲-۴-۵- ویژگی‌های سراسری به عنوان اولیه‌های صحنه [۱۴]

توصیفگرهای مهم گیبسون[2] زیر در سال ۱۹۷۹ از محیط‌های طبیعی برگرفته از ساختار سطح صحنه و تغییر این ساختارها نسبت به زمان (یا تغییرناپذیری) هستند. این جنبه‌ها مستقیماً عملیات ممکن یا تهیه‌ی مکان را کنترل می‌کنند. بنابراین، ویژگی‌های سراسری برای گرفتن اطلاعات از این سه سطح توصیف سطح صحنه، به نام‌های ساختار، تغییرناپذیری و عملیات انتخاب شده‌اند.

مجموعه‌ی هفت ویژگی برای بازتاب جنبه‌های ساختار صحنه (عمق میانگین، باز بودن، وسعت[3])، تغییرناپذیری صحنه (ناپایداری و درجه‌ی حرارت)، عملیات صحنه (پنهان‌سازی و قابلیت جهت‌یابی) انتخاب شده‌اند. به‌طور ضروری، مجموعه‌ی ویژگی‌های سراسری که اینجا فهرست شده، میانگین جامع[4] نیست. همانطور که ویژگی‌های دیگر مانند ذات طبیعی یا ناهمواری[5] بافت و تعداد و تنوع سطوح صحنه) نشان داده‌اند، توصیفگرهای مهمی از محتویات صحنه باشند. بلکه اینجا هدف به‌دست آوردن مقداری پراکندگی در این‌که صحنه‌های دنیای واقعی چگونه در ساختار، تغییرناپذیری و عملیات تغییرکرده و آزمایش این‌که کدامیک از این اطلاعات در بازنمایی صحنه‌های طبیعی نقش دارد، است.

۲-۴-۵-۱- ویژگی‌های ساختار صحنه

فعالیت محاسباتی قبل نشان داده است که دسته‌های پایه- سطح صحنه طبیعی تمایل به داشتن ساختار مکانی خاص (یا پوشش مکانی) که به خوبی از ویژگی‌های عمق میانگین، باز بودن و وسعت به دست می‌آیند، دارند. به طور مختصر، ویژگی سراسری عمق میانگین مطابق با مقیاس یا اندازه فضای شامل صحنه در محدوده یک زاویه دید نزدیک به محیط وسیع است . درجه باز بودن اندازه محفظه پوششی را نشان می‌دهد؛ درحالی‌که درجه وسعت به چشم‌انداز طرح مکانی صحنه باز می‌گردد. برای مثال، یک صحنه «مسیر در میان جنگل» ممکن است با این ویژگی‌ها مثلاً «یک محیط با تعدیل عمق و چشم‌انداز قابل ملاحظه»

---

[1] forest before trees
[2] Gibson
[3] expantion
[4] exhustive
[5] roughness

نمایش داده شود. بنابراین، این ویژگی‌های مکانی ممکن است مستقیماً از تصاویر با استفاده از ویژگی‌های سطح پایین تصویر به طور نسبی محاسبه شوند.

۲-۴-۵-۲- ویژگی‌های تغییرناپذیر صحنه

درجه‌ی تغییرناپذیری صحنه یک خصوصیت ضروری سطح طبیعی است. ویژگی‌های سراسری تغییرناپذیری توصیف می‌کنند چه مقدار و با چه سرعتی نسبت به زمان سطوح تغییر می‌کنند. اینجا، ما نقش دو ویژگی از تغییرناپذیری صحنه را ارزیابی کردیم: ناپایداری و درجه حرارت. یا احتمال تغییر سطح از یک نگاه به دیگری به طور متناوب قراردارد.

ناپایداری:

نرخی را که در آن تغییرات سطح صحنه رخ می‌دهد، شرح می‌دهد مکان‌هایی با بالاترین ناپایداری می‌توانند حرکت واقعی هم‌چون یک طوفان یا آبشار را نشان دهند. مکان‌هایی با کمترین ناپایداری می‌توانند فقط تغییر در زمان زمین‌شناسی همچون یک بی‌ثمر را نشان دهند. اگر چه درک ناپایداری ممکن است از نظر طبیعی بیشتر دریک فیلم نسبت به یک تصویر ایستا مطالعه شده، انسان‌ها می‌توانند به سادگی حرکت ضمنی را از تصاویر ایستا تشخیص دهند و درواقع این حرکت ضمن نواحی یکسان مغز به عنوان حرکت پیوسته فعال می‌کند.

درجه حرارت:

این ویژگی تفاوت در شکل ظاهری بصری یک مکان در طول تغییرات روزانه و فصلی از گرما روزانه شدید یک بیابان یا یک کوه برفی خیلی سرد را نشان می‌دهد.

۲-۴-۵-۳- ویژگی‌های عملیاتی صحنه

ساختار سطوح صحنه و تغییر آن‌ها در طول زمان ترتیب اعمالی را که یک شخص می‌تواند در یک محیط انجام دهد، کنترل می‌کند. ویژگی‌های سراسری قابلیت جهت‌یابی و پنهان‌سازی مستقیماً دو نوع از تعاملات انسان- محیط را که برای درک صحنه طبیعی از فعالیت قبلی، مهم فرض شده، اندازه‌گیری می‌کند. بنابراین همان‌طور که درک انسان شامل عمل هدف گرا[1] در محیط است، تخمین بصری سریع مسیرهای امن ممکن در یک محیط برای بازماندن ضروری است. همچنین، توانایی هدایت جستجو برای چیزهایی که با محیط پوشانده شده‌اند یا می‌تواند خود را در محیط پنهان کند مقدار بقای بالا دارد.

---

[1] object-directed

۲-۴-۶- بازشناسی صحنه مقیاس بالا- مجموعه داده SUN

اگرچه پژوهش درک صحنه با حوزه‌ی محدوده‌ای از مجموعه داده‌های فعلی که تنوع کامل دسته‌ها را دربرنمی‌گیرد، محدود شده است. درحالی‌که مجموعه داده‌های استاندارد برای دسته‌بندی شیء شامل صدها کلاس مختلف از اشیاست، بزرگترین مجموعه داده ممکن ازدسته‌های صحنه تنها شامل ۱۵ کلاس است. در این مقاله، پایگاه داده درک صحنه گسترده‌ای که شامل ۸۹۹ دسته و ۱۳۰۵۱۹ تصویر است مطرح شده است. از ۳۹۷ دسته نمونه خوب برای ارزیابی الگوریتم‌های پیشرفته‌ی متعدد برای تشخیص صحنه استفاده شده و مرزهای جدید از کارایی را ساخته شده است. به علاوه، یک بازنمایی صحنه پالایشگاه -دانه[۱] برای آشکارسازی صحنه‌های جداسازی شده درصحنه‌های بزرگ را مطالعه شده است.

درحالی‌که زمینه‌های بینایی ماشین در تشخیص چندین مجموعه داده برای سازماندهی دانش حول دسته‌های شیء (۱۵۹۲۸) توسعه یافته است، یک مجموعه داده‌ی جامع صحنه‌های دنیای واقعی بطور جاری وجود ندارد (بزرگترین مجموعه داده ممکن است ازدسته‌های صحنه تنها شامل ۱۵کلاس است). منظور از صحنه، مکانی است که یک انسان می‌تواند با آن عمل کند یا مکانی که یک انسان می‌تواند جهت‌یابی کند. چه تعداد نوع صحنه وجود ندارد؟ چگونه دانش حول صحنه‌های محیط سازماندهی شده است؟ چگونه مدل‌های صحنه پیشرفته امروزی روی محیط‌های واقعی و کنترل شده اجرا می‌شوند و چگونه با کارایی انسان مقایسه می‌شود؟

تا به امروز، کار محاسباتی روی تشخیص مکان و صحنه تصاویر طبیعی را با تعداد محدودی ازدسته‌های معنایی، بازنمایی معمولی تنظیمات داخلی و خارجی کلاس‌بندی شده است. اگرچه، مجموعه محدودی از دسته‌ها درگرفتن محیط‌های فنی و متنوع که فعالیت‌های روزانه ما را تشکیل می‌دهد، شکست می‌خورد. هم‌چون اشیاء، صحنه‌ها با عملیات و رفتارهای خاص مانند غذا خوردن در یک رستوران، مطالعه در یک کتابخانه، خوابیدن در اتاق خواب مشارکت کرده‌اند. صحنه‌ها و عملکرد وابسته‌ی آن‌ها، بسیار مرتبط با ویژگی‌های بصری که فضا را می‌سازند، هستند. عملکرد محیط‌ها می‌تواند با شکل و اندازه آن‌ها (یک راهروی باریک برای راه رفتن، یک صحنه‌ی وسیع برای رویدادهای عمومی است)، اجزای سازنده آن‌ها (برف، علف، آب، چوب) یا اشیای درون آن‌ها (میز و صندلی‌ها، تجهیزات آزمایشگاهی) تعریف شده باشد.

---

[۱] finer-grained

۲-۵- جمع‌بندی

هدف اصلی پژوهش‌های بینایی محاسباتی توسعه سیستم‌های ادراک صحنه خودکار برای ایجاد و تفسیر صحنه‌های دنیای واقعی اطراف از ورودی‌های بصری ارائه شده در تصاویر دو بعدی یا قاب‌های متوالی ویدئوها است. تا به امروز، رویکردهای محاسباتی زیادی در بینایی ماشین برای دنبال کردن مسائل و چالش‌های ادراک صحنه ارائه شده است. در میان همه این کارهای بینایی در ادراک صحنه (قطعه‌بندی تصویر، آشکارسازی شیء، بازسازی سه بعدی و سطح، استنباط هم‌پوشانی‌ها و مرزها، کلاس‌بندی صحنه، تخمین زاویه دید و عمق و غیره) کلاس‌بندی صحنه نقش مهمی در ادراک صحنه با کمک بازنمایی بالاترین سطح مفهومی را که یک تصویر صحنه به تصویر می‌کشد (داخلی درمقابل خارجی، طبیعی درمقابل مصنوعی، خیابان، شهر، اداره و غیره) ایفا می‌کند.

در جدول ۲-۱ بسیاری از روش‌های بازشناسی صحنه به صورت رویکردهای بینایی محاسباتی در مقابل رویکردهای شناخت بصری، از دیدگاه سطح ویژگی: سطح پایین در مقابل سطح مفهومی و مقیاس پردازش: محلی در مقابل سراسری (کلی)، دسته‌بندی شده‌اند.

جدول ۲-۱ دسته‌بندی رویکردهای محاسباتی و شناختی برای کلاس‌بندی صحنه.

ویژگی / روش			سطح ویژگی‌ها: سطح پایین	سطح ویژگی‌ها: سطح مفهومی
رویکردهای مبتنی‌بر بینایی محاسباتی	چند وجهی (سطح ترکیب)	سطح وظایف		چارچوب کلی با استفاده از تصاویر ذهنی (نگاشت ثبت شده‌ای که یک مشخصه از صحنه را توصیف می‌کند.) [۳۶]
		سطح ویژگی	رویکرد یادگیری تولیدکننده/ تفکیک‌کننده دو فازی با استفاده از انواع ویژگی چندتایی [۳۲] لغت‌نامه بصری با آموزش یک نگاشت خود-سازماندهی شده روی بردارهای ویژگی نواحی قطعه قطعه شده تصویر [۲۸] مدل کردن وابستگی‌های مکانی محلی داده روی هیستوگرام چند مقیاسه روی جهت‌های گرادیان [۲۹]	ایجاد یک شبکه بیزین با استفاده از ویژگی‌های سطح پایین و معنایی [۲۲]
	تک وجهی (مقیاس پردازش)	سراسری	بازنمایی سراسری براساس تکستون [۳۳] الگوریتم یادگیری فعال ماشین بردار پشتیبان براساس رنگ و بافت برای استنتاج بازخورد ارتباط موثر [۳۴] بازنمایی سطح پایین سراسری تصویر براساس رنگ و لبه‌ها [۳۵]	
		محلی	سیفت [۳۹] بلاب ورلد [۲۴] هرم قابل هدایت چند مقیاسه [۲۵] بازنمایی با بردار ویژگی با ابعاد بالا به‌دست آمده از خروجی درختی از فیلترهای غیرخطی براساس شباهت بین نواحی با ویژگی‌های بافتی، مکانی یا رنگی خاص رویکرد مشابه بازنمایی اختلافات [۲۱] استخراج ویژگی بافت برای نواحی شکل-دلخواه [۱۹]	بازنمایی به صورت مجموعه‌ای از وصله‌های محلی که به طور خودکار روی نقاط تغییرناپذیر نسبت به مقیاس [۲۶] ماتریس هم‌رخدادی سطح خاکستری [۲۴] بازنمایی روابط جهتی بین اشیا با رشته دو بعدی و متغیرهای آن [۳۰] مدل‌کردن صحنه‌های بصری در مجموعه‌های تصویر، براساس ویژگی‌های غیرقابل‌تغییر محلی و pLSA [۲۰] بازنمایی هرم مکانی تصویر با ساخت یک هسته انطباق هرمی برای یافتن یک انطباق تخمینی بین دو مجموعه از عناصر استفاده شده [۲۷] دو مدل جدید ABSpLSA و FSI-pLSA [۲۳]
رویکردهای الهام گرفته از شناخت بصری		سراسری		مدل کردن شکل صحنه (بازنمایی سراسری از پوشش مکانی) [۲] توصیف صحنه- محور از ویژگی‌های پوشش مکانی [۳۷] ساخت gist صحنه [۲۹] آماره‌های دسته صحنه‌های طبیعی [۴۸] تخمین عمق از ساختار تصویر[۳۴]

فصل سوم: تئوری

۳-۱- مقدمه

در این فصل به بررسی تئوری دو روش موردنیاز برای پیاده‌سازی الگوریتم پیشنهادی می‌پردازیم. همان-طور که در فصول قبل اشاره شد و در فصل ۴ هم به تفصیل تشریح می‌شود، الگوریتم پیشنهادی براساس روش‌های مبتنی بر ویژگی محلی سیفت و ویژگی سراسری جیست نرمال شده می‌باشد که در ادامه توضیحاتی در مورد این دو روش آورده شده است:

۳-۲- روش‌های مبتنی بر ویژگی محلی سیفت [۴۰]

تبدیل ویژگی غیرقابل تغییر با مقیاس (سیفت) ابتدا توسط لووه[۱] به عنوان ترکیبی از توصیفگر ناحیه دلخواه DoG و یک تصویر ویژگی انطباق معرفی شده بود. اگرچه هر دو مؤلفه به صورت ایزوله استفاده شده‌اند؛ به خصوص، دنباله‌ای از مطالعات تایید کرده که توصیفگر سیفت برای ترکیب با همه آشکارسازهای ناحیه که بالا اشاره شد، مناسب است.

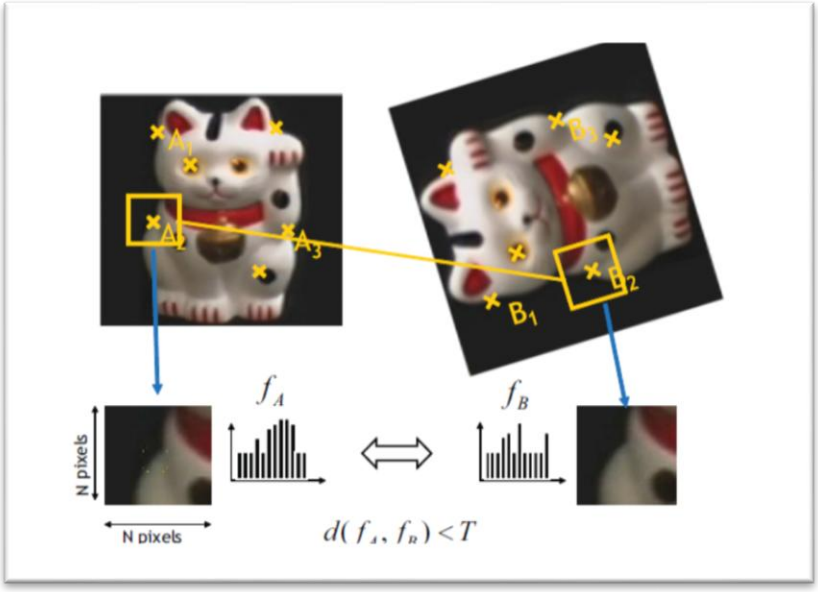
توصیفگر سیفت کمک می‌کند تا قدرت برای تغییرات روشنایی و انتقالات مکانی کوچک با کدگذاری اطلاعات تصویر در یک مجموعه محلی شده از هیستوگرام‌های جهت‌دار، به‌دست می‌آید. محاسبه توصیفگر از یک ناحیه نرمال شده مقیاس و چرخش استخراج شده با یکی از آشکارسازهای اشاره شده شروع می‌شود. به عنوان گام اول، اندازه و جهت گرادیان تصویر، حول موقعیت نقطه کلیدی با استفاده از مقیاس ناحیه برای انتخاب سطح محو گاوسی (سطح هرم گاوسی که این محاسبه اجرا شده است) نمونه‌برداری شده است. نمونه‌برداری در یک شبکه منظم ۱۶×۱۶ که ناحیه مورد علاقه را می‌پوشاند، اجرا شده است. برای هر موقعیت نمونه‌برداری شده، جهت گرادیان در یک شبکه ۴×۴ از هیستوگرام‌های جهت گرادیان با ۸ جهت اولیه در هرکدام که با انطباق اندازه گرادیان پیکسل و با یک تابع وزن‌دهی گاوسی با نصف اندازه ناحیه، وزن‌دهی شده است. هدف این پنجره گاوسی بالا بردن وزن پیکسل‌های به میانه ناحیه است که کمتر تحت تاثیر بی‌دقتی‌های موقعیت‌یابی کوچک آشکارساز ناحیه مورد علاقه قرار گرفته است.

این رویه برای یک پنجره ۲×۲ کوچکتر در شکل ۳-۱ نشان داده شده است. انگیزه انتخاب این بازنمایی این است که تظریف مکانی سخت، برای جابجایی‌های کوچک بخاطر خطاهای ثبت بدون تاثیر بیش از حد

---

[۱] lowe

آشکارساز مجاز می‌شود. در همان زمان، بازنمایی با ابعاد بالا، قدرت تفکیک‌پذیری کافی برای تمییز قابل اعتماد تعداد زیادی از نقاط کلیدی را فراهم می‌کند.



شکل ۳-۱ رویه به‌دست آوردن توصیفگر سیفت برای یک پنجره ۲×۲ در تصویر.

هنگام محاسبه توصیفگر، جلوگیری از همه اثرات مرزی با توجه به انتقالات فضایی و تغییرات جهت کوچک ضروری است. بنابراین، هنگامی که اطلاعات گرادیان پیکسل نمونه‌برداری شده به هیستوگرام مکان/ جهت سه بعدی وارد می‌شود، توزیع آن باید به صورت هموار در میان بخش‌های هیستوگرام مجاور با استفاده از درون‌یابی سه خطی صورت گیرد.

هنگامی که همه درایه‌های هیستوگرام جهت کامل شده‌اند، درایه‌ها به شکل یک بردار ویژگی ۴×۴×۸=۱۲۸ بعدی الحاق شده‌اند. نرمال کردن نهایی روشنایی، رویه استخراج را کامل می‌کند. برای این هدف، بردار ابتدا به یک طول واحد نرمال شده پس برای تغییر تضاد تصویر سازگار می‌شود. بنابراین همه ابعاد آینده با یک مقدار بیشینه ۰/۲ آستانه‌گذاری شده و بردار دوباره به طول واحد نرمال می‌شود. این گام آخر برای تنظیم تغییرات روشنایی غیرخطی به دلیل اشباع دوربین یا اثرات مشابه است.

۳-۳- روش‌های مبتنی بر ویژگی سراسری جیست [۱، ۱۰ و ۱۸]

۳-۳-۱- جیست دو بعدی صحنه

جیست دو بعدی صحنه ویژگی مفهومی دیگری است که از آماره‌های سراسری[۱] یک تصویر برای به‌دست آوردن «جیست» یا «قاب مفهوم[۲]» تجربه بصری با استفاده از ویژگی‌های آماری صحنه به جای اشیا یا نواحی محلی استفاده می‌کند. اولیوا و تورالبا یک مدل محاسباتی برای بازشناسی صحنه‌های دنیای واقعی (چهار صحنه طبیعی و چهار صحنه مصنوعی) که قطعه قطعه‌سازی و پردازش اشیا با نواحی خاص را انجام می‌دهند، را مطرح کردند. رویه براساس یک بازنمایی بسیار سطح پایین از صحنه است که به آن پوشش مکانی[۳] می‌گویند. این بازنمایی شامل خصوصیات ادارکی است: طبیعی بودن[۴] (در مقابل مصنوعی)، باز بودن[۵] (حضور یک خط افقی)، سختی[۶] (پیچیدگی انحرافی)، وسعت[۷] (چشم‌انداز در صحنه‌های مصنوعی) و ناهمواری (انحراف از افق در تصاویر طبیعی). اگرچه، توزیع هر ویژگی نمی‌تواند به شکلی که هستند، درک شود و مهم‌تر این‌که، نمی‌توانند مستقیماً برای ناظران انسانی معنی‌دار باشند؛ هر ویژگی منطبق با یک بعد در فضای پوشش مکانی است و دیگر ساختار مکانی اصلی یک صحنه را بازنمایی می‌کند. بنابراین، این ویژگی‌ها نشان می‌دهند که این ابعاد ممکن است به طور قابل اعتمادی با استفاده از اطلاعات محلی شده طیفی و دقیق تخمین زده شوند. این مدل، یک فضای چند بعدی را مدل می‌کند که در صحنه‌هایی که عضویت را در دسته‌هایی که نزدیک به هم نگاشت شده‌اند، به اشتراک می‌گذارند. بنابراین انتساب یک تفسیر خاص به هر بعد ممکن است: برای ویژگی باز بودن، تصویر به یک محیط باز یا بسته نسبت داده می‌شود و غیره.

مطالعات در ادراک صفحه نشان داده است که ناظران در صحنه دنیای واقعی در یک نگاه تشخیص می‌دهند. در طی این فرآیند ضروری دیدن، سیستم بینایی یک بازنمایی مکانی از دنیای بیرونی که به اندازه کافی برای فهمیدن معنای صحنه، تشخیص تعداد شی و دیگر اطلاعات برجسته[۸] تصویر غنی است را شکل می‌دهد. این بازنمایی را جیست یک صحنه می‌گویند که شامل همه سطوح پردازش، از ویژگی‌های سطح پایین (مانند رنگ، فرکانس‌های مکانی) تا ویژگی‌های میانی تصویر (مانند مساحت و حجم) و اطلاعات سطح

---

۱ global
۲ Concept frame
۳ spatial envelope
۴ naturalness
۵ openness
۶ roughness
۷ expansion
۸ salient

بالا (مانند اشیاء، فعال‌سازی دانش‌سازی) است. بنابراین، می‌تواند در دو سطح ادارکی[۱] و مفهومی[۲] مطالعه شود.

۳-۳-۲- تعریف «جیست یک صحنه»

با یک نگاه گذرا به هر صحنه پیچیده دنیای واقعی، یک ناظر می‌تواند تنوعی از اطلاعات مفهومی و معنایی[۳] را درک کند. تجربه درک هر چیز در نگاه اول، باتوجه به پیچیدگی بصری صحنه، هنگام تماشای تلویزیون و تعویض سریع کانال‌ها تجربه شده باشد: فقط با یک نگاه آنی هر تصویر، بیننده می‌تواند معنای آن را (یک سیاست‌مدار، یک مسابقه اتومبیلرانی، اخبار، کارتون، غیره) مستقل از پارازیت و تنوع جزئیات درک کند. این تعریف به جیست یک صحنه برمی‌گردد.

مطالعات رفتاری نشان داده که بینندگان می‌توانند دسته‌بندی سطح پایه صحنه (مثلاً یک خیابان) و طرح کلی مکانی آن (مثلاً یک خیابان با بلوک های عمودی بلند در هر دو طرف) را به خوبی دیگر اطلاعات ساختار کلی (مانند حجم زیاد در دیدگاه) در کمتر از ۱۰۰ میلی ثانیه تشخیص دهند. ناظران همچنین ممکن است تعداد کمی از اشیاء (مانند اتومبیل قرمز و اتومبیل سبز)، محتوایی که در آن به نظر می‌رسد (مثلاً در خیابان پارک شده است) و دیگر مشخصه‌های سطح پایین نواحی که به‌خصوص برجسته هستند را به خاطر بسپارند.

هم‌زمان با توسعه‌ی جیست، فعال‌سازی خودکار یک چارچوب از اطلاعاتی معنایی شامل متون (مثلاً اعمالی که در یک صحنه رخ داده و دانش مرتبط به صحنه (مثلاً معمولاً شبیه به یک صحنه خاص است) و پیش‌بینی‌های اشیایی که احتمالاً در محیط یافت می‌شوند، وجود دارد.

۳-۳-۳- ماهیت جیست

چون جیست شامل همه سطوح اطلاعات بصری- از ویژگی‌های سطح پایین (مانند رنگ) تا اطلاعات سطح بالا (مانند فعال‌سازی دانش معنایی) است، می‌تواند در دو سطح ادراکی و مفهومی شامل اطلاعات معنایی باشد که هنگام مشاهده یک صحنه یا بلافاصله بعد از این‌که صفحه ناپدید شد نتیجه می‌شوند. جیست، مفهومی غنی‌سازی شده و به عنوان اطلاعات مفهومی برگرفته از مراحل اولیه پردازش دانش بصری، اصطلاح شده است.

---

[۱] perceptual
[۲] conceptual
[۳] semantic

۳-۳-۴- جیست مفهومی

فعالیت‌های اخیر مری پاتر[۱] اطلاعات مفهومی را که ناظران می‌توانند به سرعت از یک تصویر درک کنند، شرح داده است. در مطالعه اصلی، پاتر و لوی به ناظران اجازه می‌دهند یک نگاه آنی به دنباله‌ای از تصاویر معنادار قبل از آزمایش حافظه آن‌ها از این تصاویر داشته باشند. هنگامی که هر تصویر به تنهایی ۱۰۰ میلی‌ثانیه ظاهر می‌شود، به راحتی به‌خاطر سپرده می‌شود. اگرچه، هنگامی که دریک الگوی بازنمایی بصری دنباله‌ای سریع[۲] تعبیه شدند، کارایی‌ها کاهش یافتند. در مطالعه دوم در سال ۱۹۷۶ ناظران قبل از ظهور ممکن یک تصویر در دنباله RSVP قرار گرفتند. نشانه، یک تصویر یا یک توصیف کلامی کوتاه از تصویر مثل عبارت «یک گردش در ساحل» بوده و درخواست آشکارسازی آن شده است. نتایج به خوبی، با کارایی آشکارسازی به ترتیب ۶۰ درصد و ۸۰ درصد در سرعت های ۱۲۵ و ۱۵۰ میلی ثانیه در هر تصویر است. همان‌طور با ۱۲ و ۳۰ درصد در مرحله یادآوری تشخیص کنترل، بهبود یافت. با دیگر شواهد تجربی، نتایج به‌دست آمده توسط مری پاتر نشان داد که در طول یک RSVP، پردازش یک تصویر جدید ممکن است پیوستگی در حافظه کوتاه‌مدت تصویر قبلی را به هم بزند. با ۱۰۰ میلی ثانیه هر تصویر فوراً درک می‌شود و بنابراین به نظر می‌رسد ناظران حجم زیادی از اطلاعات بصری را درک کرده‌اند، اما یک تاخیر در هزاران میلی‌ثانیه برای تثبیت تصویر در حافظه نیاز است. هنگامی که تثبیت شد، جیست مفهومی می‌تواند به عنوان یک توصیفی کلامی از تصویر صحنه که شامل آنچه درک و استنباط شده بود، بازنمایی شود.

۳-۳-۵- جیست ادراکی

حصول اطمینان از محتوای ادراکی جیست شامل تعیین ویژگی‌هایی از تصویر (مثلاً فرکانس مکانی، رنگ، بافت) است که یک بازنمایی ساختاری از صحنه را ایجاد می‌کنند.

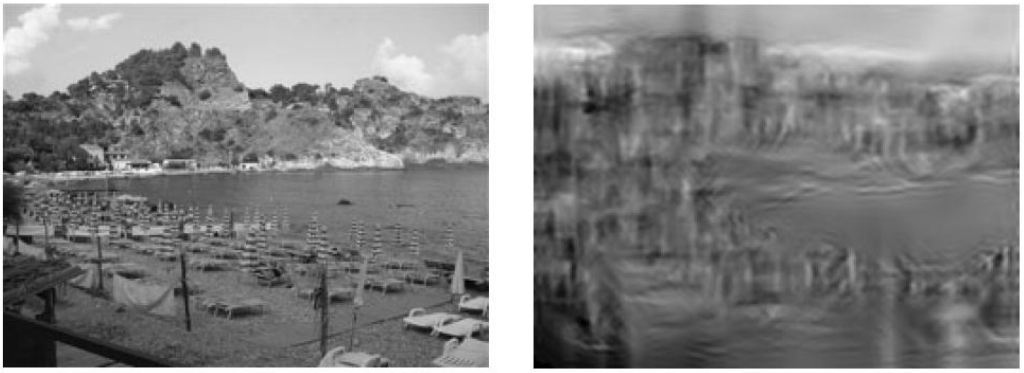
با دست‌کاری میزان در دسترس بودن این ویژگی‌های تصویر (مثلاً فیلتر کردن لبه‌های یک تصویر) مانند محدودیت‌های تحمیل شده می‌توان از آن تجربه اطلاعات موردنظر برای ساخت یک جیست مفهومی ساختاری را تعیین کرد و بنابراین شناسایی صحنه ممکن می‌شود.

الیوا و اسچینز[۳] در سال‌های ۱۹۹۴ تا ۲۰۰۰ نشان دادند که یک توصیف سخت[۴] از صحنه ورودی (جابجایی جهت‌دار در ترتیب مکانی خاص در وضوحی به کمی ۴ دور در هر تصویر) ممکن است تشخیص

---

۱ Mary Potter
۲ rapid serial visual presentation
۳ Schyns
۴ coarse

را قبل از شناسایی اشیای پردازش شده آغاز کند. به‌طور مشابه ممکن است، ساختار یک صحنه به سرعت از ترکیب یک طرح کلی از بخش‌ها (ترتیب مکانی شکل‌های حجمی ساده مانند جئون‌ها[1]) با یک تصویر طرح سخت از تراکم بافت به‌دست آید. در شکل ۳-۲ یک بازنمایی جیست صحنه که اطلاعات ساختاری کافی برای استنتاج دسته صحیح صحنه را حفظ می‌کند، نشان داده شده است.



شکل ۳-۲ نمایش یک بازنمایی جیست صحنه که اطلاعات ساختاری کافی برای استنتاج دسته صحیح صحنه را حفظ می‌کند [۱۰].

۳-۴- جمع‌بندی

در این فصل توضیحاتی در مورد روش‌های سیفت و جیست که اجزای تشکیل‌دهنده روش پیشنهادی هستند، آورده‌ایم. در فصل بعد جزئیات کاربرد آن‌ها تشریح می‌شود.

[1] genos

فصل چهارم: بررسی و تحلیل الگوریتم پیشنهادی

۴-۱- مقدمه

در این فصل به بررسی الگوریتم پیشنهادی می‌پردازیم. این الگوریتم سعی بر استخراج ویژگی‌هایی دارد تا براساس آن‌ها تصاویر ورودی را تا حد امکان با دقت بیشتری نسبت به روش‌های موجود و در کمترین زمان ممکن دسته‌بندی کند و نسبت به تغییرات مقیاس، چرخش و روشنایی پایدار باشند. با توجه به مطالب مطرح شده در فصول گذشته چند اشکال اساسی در روش‌های مبتنی بر ویژگی‌های سطح پایین وجود دارد.

در ادامه این فصل ابتدا به بررسی چالش‌هایی که در الگوریتم‌های پیشین که به تفصیل در فصل دوم بیان شد، می‌پردازیم و دلایل لزوم الگوریتم پیشنهادی را بررسی می‌کنیم. سپس روشی جدید ارئه می‌دهیم تا علاوه بر داشتن مزایای این روش‌ها، اشکالات آن‌ها را نیز نداشته باشند. سپس الگوریتم را به‌صورت مرحله به مرحله تحلیل و بررسی می‌کنیم و در بخش آخر به بررسی الگوریتم پیشنهادی می‌پردازیم.

۴-۲- چالش‌های روش‌های قبلی

یک محدودیت بازنمایی‌های کلی ارائه شده نیاز به این است که الگوهای دو بعدی باید خیلی مرتب باشند. این مسئله برای اطمینان از زمان ایجاد یک مجموعه آموزشی برچسب‌دار از تصاویر، پرهزینه است. برای مثال، اگر یک کلاس چهره با لیست‌هایی مرتب از شدت نورهای نمونه‌های چهره تمام رخ نمایش داده شود، حتی یک جابجایی کوچک در یک نمونه باعث می‌شود پیکسل‌های بینی در سمت راست بلند نبوده و بازنمایی کلی شکست می‌خورد.

یک هیستوگرام رنگی، فرکانس را همراه با این‌که هر رنگ در میان همه پیکسل‌های پنجره چقدر ظاهر می‌شود اندازه‌گیری می‌کند. بدون درنظرگرفتن ترتیب، هیستوگرام، تغییرناپذیری شرایط زاویه دید و بعضی تغییرات هم‌پوشانی جزئی را پیشنهاد می‌کند. این حساسیت باعث می‌شود امکان استفاده از تعداد کمی از زوایای دید برای بازنمایی یک شی با فرض یک فضای جهان بسته با اشیای رنگی متمایز به‌وجود آید. اولین رویکرد، استفاده از هیستوگرام‌های رنگی برای آشکارسازی شی خاص توسط سویین[۱] و بلارد[۲] مطرح شد. این دیده توسط شیله[۳] و کراولی[۴] برای هیستوگرام‌های چند بعدی عملکرهای همسایه تصویر محلی در محدوده مشتقات ساده تا ترکیب مشتقات پیچیده‌تر استخراج شده در چندین مقیاس، گسترش یافت. آن

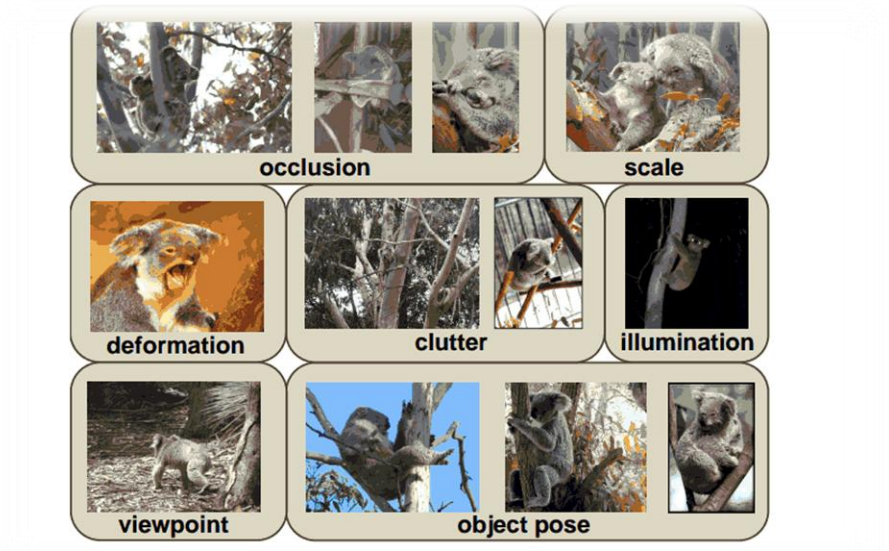
---

[۱] swin
[۲] ballard
[۳] schiele
[۴] crowely

هیستوگرام‌ها به طور موثر، توزیع احتمال ترکیب ویژگی خاص در شیء را در نظر گرفتند. علاوه بر رویکرد آشکارسازی کلی با انطباق هیستوگرام های کامل، شیله و کراولی همچنین یک روش محلی برای تخمین بیشینه‌ی احتمال پسین اشیای مدل مختلف از تعداد کمی همسایه نمونه در تصویر آزمایشی مطرح کردند. بازنمایی‌های کلی که در بالا شرح داده شد منجر به رویکرد آشکارسازی بر اساس مقایسه‌ی تصاویر کلی یا پنجره‌های تصویر کلی شدند. این رویکردها برای یادگیری ساختار کلی شیء مناسب هستند اما نمی‌توانند از عهده هم‌پوشانی‌های جزئی، تغییرات شدید زاویه دید یا اشیای تغییر شکل داده شده برآیند.

همچنین روش‌های قبلی که از ترکیب ویژگی‌های سراسری و محلی استفاده کرده‌اند، بردارهای ویژگی استخراج شده از آن‌ها بزرگ بوده در نتیجه محاسبه این بردارها از نظر حافظه پرمصرف و زمان‌بر است. یکی از اهداف ما، استفاده از الگوریتم‌های مناسب برای کاهش بردار ویژگی‌های استخراج شده برای فرستادن به کلاس‌بند است. باتوجه به اینکه روش‌های مبتنی‌بر ویژگی‌های سراسری از ساختار کلی تصویر بهره می‌گیرند برای دسته صحنه‌های با ساختار پیچیده و جزئیات زیاد مانند ساختمان بلند به خوبی جواب نمی‌دهند. به علاوه، روش‌های مبتنی‌بر ویژگی‌های محلی هم با توجه به رویکرد قطعه‌بندی آن‌ها نتیجه مطلوبی برای بازشناسی دسته صحنه‌هایی مانند منظره باز که طرح مکانی ساده‌ای دارند، به‌دست نمی‌آورند.

در شکل ۴-۱ برخی از دشواری‌هایی که معمولاً هنگام دسته‌بندی تصاویر با آن‌ها روبه‌رو می‌شویم، نشان داده شده است [۸].



شکل ۴-۱ برخی از دشواری‌هایی که معمولاً در دسته‌بندی تصاویر با آن‌ها روبه‌رو هستیم [۸].

۴-۳- رویکرد کلی الگوریتم پیشنهادی

در این بخش به بیان رویکرد کلی الگوریتم پیشنهادی می‌پردازیم. ورودی این الگوریتم تصاویر صحنه‌های ثبت شده از محیط‌های خارجی می‌باشد که یک ماتریس از تصویر رنگی با ابعاد ۳×۲۵۶×۲۵۶ است، سپس بردارهای جیست و سیفت آن را به صورت مجزا به‌دست می‌آوریم که این مرحله استخراج ویژگی است. بردارهای ویژگی به‌دست آمده شامل مقادیر جیست و سیفت ماتریس تصویر ورودی است. هدف این مرحله، استخراج ویژگی‌هایی است که ب اساس آن تصاویر را با استفاده از این ویژگی‌های جدید توصیف کنیم. پس از استخراج این ویژگی‌ها و تعریف فضای جدید ویژگی‌ها، براساس آن‌ها ابتدا تصاویر را به صورت مجزا دسته‌بندی می‌کنیم و سپس از ترکیب نتایج دسته‌بندی دو روش با استفاده از قانون ترکیبی ضرب[۱] برای به‌دست آوردن نتایج نهایی استفاده می‌کنیم که نحوه انجام این کار در بخش بعدی به تفصیل توضیح داده می‌شود.

تصاویر مجموعه داده باید در ۸ پوشه مجزا باشند که شماره هر پوشه، نشان‌دهنده کلاس آن دسته تصاویر است. پس از به‌دست آوردن بردار ویژگی‌ها با استفاده از یک کلاس‌بند ماشین بردار پشتیبان با هسته گاوسی برای دسته‌بندی براساس جیست و با هسته اشتراک هیستوگرام[۲] برای دسته‌بندی براساس ویژگی سیفت را آموزش می‌دهیم و سپس به آزمایش می‌پردازیم. در مرحله پایانی، نتایج به‌دست آمده از دو کلاسبند با یک قانون ترکیبی برای ایجاد کلاس نهایی ترکیب می‌شوند.

۴-۴- بررسی مرحله به مرحله الگوریتم پیشنهادی

در این بخش به تفصیل به بررسی الگوریتم پیشنهادی می‌پردازیم. همانطور که در فصل قبل اشاره شد ساختار پایه‌ی استفاده شده در این پایان‌نامه، ساختار استاندارد الگوریتم جیست و سیفت است. الگوریتم شامل ۴ مرحله است که در ادامه به تفصیل توضیح داده شده‌اند. این الگوریتم که به دسته‌بندی صحنه‌های خارجی می‌پردازد شامل مراحل اصلی زیر می‌باشد:

۱) استخراج ویژگی جیست از ماتریس تصویر و به‌دست آوردن جیست نرمال شده
۲) استخراج ویژگی سیفت از ماتریس تصویر
۳) کلاس‌بندی با استفاده از ماشین بردار پشتیبان براساس هردو ویژگی به صورت مجزا
۴) ترکیب نتایج کلاس‌بندی با استفاده از قانون ترکیبی ضرب

---

[۱] product rule
[۲] histogram intersection kernel

مرحله اول: استخراج ویژگی جیست از ماتریس تصویر و به‌دست آوردن جیست نرمال شده

توصیفگر جیست انرژی خروجی یک بانک ۲۴ فیلتری را محاسبه می‌کند. فیلترهای به‌دست آمده، فیلترهای گابور- مانند تنظیم شده با ۸ جهت در ۴ مقیاس مختلف هستند. سپس، مجذور هر فیلتر روی یک شبکه ۴×۴ میانگین‌گیری شده است[۱۶]. برای انجام مرحله اول باید روی ماتریس تصویر، استخراج ویژگی انجام دهیم. یعنی به منظورکاهش زمان کلاس‌بندی، ابعاد ماتریس تصویر را کاهش می‌دهیم. ویژگی استخراج‌شده از تصویر جیست می‌باشد. برای به‌دست آوردن ویژگی جیست از تابع زیر استفاده می‌شود:

```
[gist, GISTparam] = LMgist (Image, GISTparam)
```

این تابع با دریافت تصویر و پارامترهای جیست، این ویژگی را محاسبه و یک بردار ویژگی ۵۱۲×۱ برمی‌گرداند که بردار ویژگی موردنظر برای فرستادن به کلاس‌بند می‌باشد. ساختارداده GISTparam از چند مولفه زیر تشکیل شده است:

```
GISTparam.imageSize = [256 256];
GISTparam.orientationsPerScale = [8 8 8 8];
GISTparam.numberBlocks = 4;
GISTparam.fc_prefilt = 4;
```

مولفه سوم نشان‌دهنده جهت‌های موردنظر در ۴ مقیاس است (از فرکانس‌های پایین تا بالا).

به‌دست آوردن جیست شامل دو مرحله پیش‌فیلتر کردن تصویر و سپس فیلتر کردن و جمع‌آوری انرژی-های خروجی است. با مقداردهی ویژگی اندازه به صورت GISTparam.imageSize = [256 256] می‌توان اندازه تصویر را تغییر داد. قابل ذکر است که الگوریتم محاسبه جیست تصویر، برای تصاویری که مربعی نیستند هم به درستی جواب می‌دهد.

مرحله دوم: استخراج ویژگی سیفت از ماتریس تصویر

برای انجام مرحله دوم عملیاتی مشابه مرحله اول انجام می‌شود؛ با این تفاوت که در این مرحله ویژگی سیفت تصویر محاسبه می‌شود. ساختار داده Vwparamsift از چند مولفه زیر تشکیل می‌شود:

```
VWparamsift.imageSize = [256 256];
VWparamsift.grid_spacing = 1;
VWparamsift.patch_size = 16;
VWparamsift.NumVisualWords = 200
VWparamsift.Mw = 2;
```

```
VWparamsift.descriptor = 'sift';
VWparamsift.w = VWparamsift.patch_size/2;
```

مولفه اول اندازه تصویر، مولفه دوم فاصله بین مراکز شبکه‌ها ، مولفه سوم اندازه وصله برای محاسبه توصیفگر سیفت (باید عاملی از ۴ باشد)، مولفه چهارم تعداد واژه‌های بصری، مولفه پنجم تعداد مقیاس‌های مکانی برای هیستوگرام هرم مکانی، مولفه ششم نوع توصیفگر و مولفه آخر مرز توصیفگر سیفت را تنظیم می‌کنند. همانند الگوریتم جیست، این الگوریتم هم برای تصاویری که مربعی نیستند به درستی اجرا می‌شود.

با استفاده از دستور زیر یک لغت‌نامه از ده تصویر اول مسیر برای مجموعه داده تهیه می‌کنیم:

```
VWparamsift = LMkmeansVisualWords (D(1:10:end), HOMEIMAGES, VWparamsift);
```

و سپس با دستوری که در ادامه آورده شده واژه بصری را برای تمام تصاویر مجموعه داده به‌دست می‌آوریم:

```
[VWsift, sptHistsift] = LMdenseVisualWords (D, HOMEIMAGES, VWparamsift);
```

خروجی این دستور ماتریس sptHistsift است که شامل مقادیر سیفت تمامی تصاویر مجموعه داده می‌باشد.

مرحله سوم: کلاس‌بندی با استفاده از ماشین بردار پشتیبان براساس هردو ویژگی به صورت مجزا

برای این مرحله یعنی کلاس‌بندی صحنه‌ها از یک ماشین بردار پشتیبان با کرنل گاوسی برای دسته‌بندی براساس جیست و یک کلاس‌بند ماشین بردار پشتیبان با هسته اشتراک هیستوگرام[1] برای دسته‌بندی براساس ویژگی سیفت استفاده می‌کنیم. قبل از شروع کلاس‌بندی با تخصیص شماره پوشه هر دسته صحنه به یک کلاس، ماتریس کلاس‌ها را تشکیل می‌دهیم. سپس تصاویر را به سه بخش آموزشی، آزمایشی و ارزیابی تقسیم می‌کنیم. پارامترهای ماشین بردار پشتیبان موردنظر، پارامتر ترتیب‌دهی، بیشترین تعداد گام‌های نیوتن و معیار توقف svm هستند. پارامترها و داده‌های آموزشی و آزمایشی را به تابع kernel() داده و درنهایت کلاس‌بندی را با تابع primalsvm() انجام می‌دهیم.

مرحله چهارم: ترکیب نتایج کلاس‌بندی با استفاده از قانون ترکیبی ضرب

در مرحله آخر با استفاده از قاعده ترکیبی ضرب، کلاس نهایی را براساس کلاس‌های خروجی حاصل از کلاس‌بندی با ویژگی‌های جیست و سیفت به‌دست می‌آوریم. به این صورت که، در کلاس نهایی برای هر

[1] histogram intersection kernel

مولفه، کلاسی که بیشترین احتمال وقوع را دارد درنظر می‌گیریم. برای کلاس‌های توزیع هم‌احتمال کلاسی به عنوان کلاس نهایی انتخاب می‌شود که شرط معادله زیر را برقرار کند:

$$\prod_{i=1}^{R} P(w_j|x_i) = max_{k=1}^{m} \prod_{i=1}^{R} P(w_k|x_i) \quad ۱-۳$$

که در این معادله m تعداد کلاس‌های ممکن ($w_1, \ldots, w_m$)، x_i بردار اندازه‌ای که با کلاس‌بند i ام استفاده شده و R تعداد کلاسبندهاست که در اینجا مقدار R برابر ۲ است [۵۰].

۴-۵- جمع‌بندی

این الگوریتم که مراحل آن در بخش قبل بیان شد، با استخراج مفاهیم سطح بالا از تصویر می‌کوشد تا همزمان از اطلاعات سراسری و محلی تصویر استفاده کند. آنچه موجب بهبود نتایج می‌شود این است که استخراج ویژگی با ابعاد کمتر از برخی روش‌های قبلی و همچنین استفاده از مزایای هردو روش سراسری و محلی به ترتیب برای دسته صحنه‌هایی که طرح کلی ساده و پیچیده‌ای دارند، انجام می‌شود. در فصل بعد، به بررسی و ارزیابی نتایج روش پیشنهادی می‌پردازیم.

فصل پنجم: پیاده‌سازی، ارزیابی کارایی و مقایسه با روش‌های دیگر

۵-۱- مقدمه

در این فصل به ارزیابی الگوریتم پیشنهادی می‌پردازیم. ابتدا مجموعه داده استفاده شده را معرفی می‌کنیم. سپس به ارزیابی استخراج ویژگی سراسری و محلی پرداخته و میزان موفقیت روش ارائه شده در این زمینه را ارزیابی می‌کنیم. نتایج را با نتایج به‌دست آمده از پیاده‌سازی برخی از الگوریتم‌های دسته‌بندی صحنه مشابه دیگر مقایسه می‌نماییم.

تمامی آزمایش‌های این پایان‌نامه با نرم‌افزار متلب نسخه 2012a در سیستم‌عامل ویندوز ۷ پیاده‌سازی و روی یک کامپیوتر ایسوس با پردازنده مرکزی ۵ هسته‌ای ۱/۸ گیگاهرتز و حافظه داخلی ۶ گیگابایت انجام شده‌اند.

۵-۲- معرفی مجموعه داده

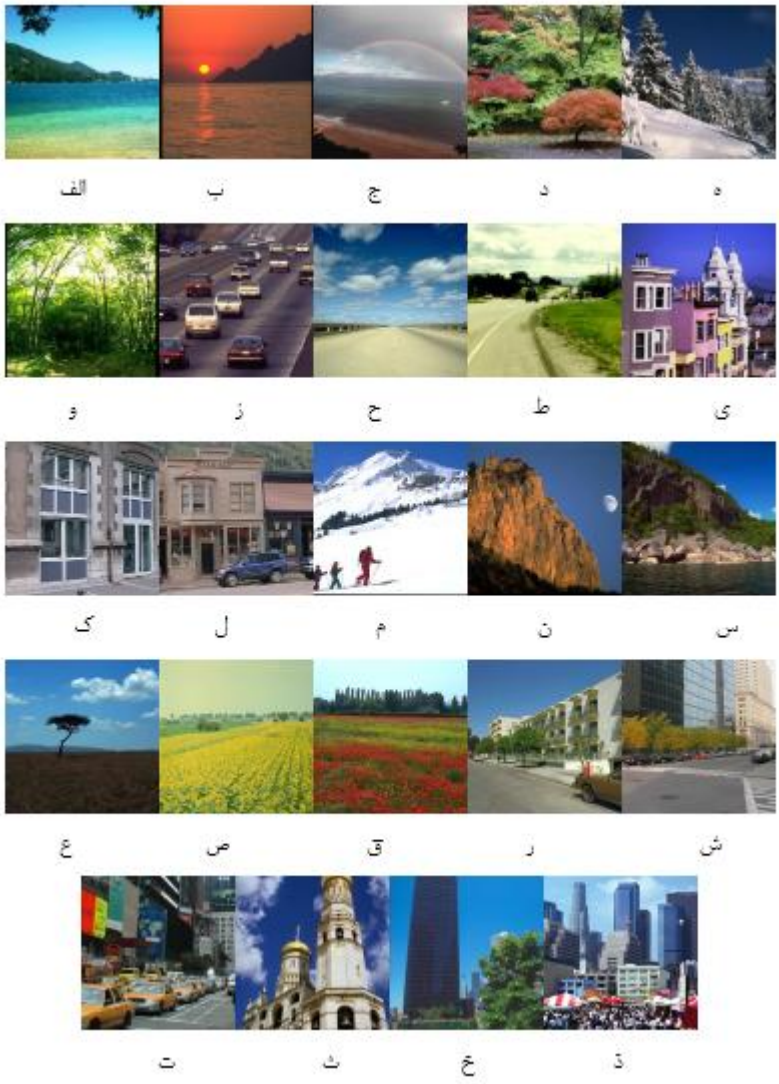
در این بخش به معرفی مجموعه داده صحنه‌های خارجی که در اکثر پژوهش‌های دسته‌بندی صحنه اخیر مورد توجه بوده است، می‌پردازیم .این مجموعه داده بسیار چالش‌برانگیز بوده وبسیاری از الگوریتم‌های شناسایی صحنه، امروزه بر روی آن آزمایش می‌شوند. این مجموعه داده، مجموعه داده‌ی صحنه‌های «۸دسته صحنه خارجی» می‌باشد که توسط اولیوا و تورالبا معرفی شده است.

این مجموعه داده شامل ۸ دسته صحنه خارجی است که تعداد تصاویر موجود در هر دسته متفاوت و در مجموع ۲۶۸۸ تصویر دارد. این مجموعه داده شامل دسته‌های ساحل[۱] (۳۶۰ تصویر)، جنگل[۲] (۳۲۸ تصویر)، بزرگراه[۳] (۲۶۰ تصویر)، درون شهر[۴] (۳۰۸ تصویر)، کوه[۵] (۳۷۴ تصویر)، منظره باز[۶] (۴۱۰ تصویر)، خیابان[۷] (۲۹۲ تصویر) و ساختمان بلند[۸] (۳۵۶ تصویر) می‌باشد. تصاویر به صورت رنگی در فضای RGB ذخیره شده‌اند. تصاویر این مجموعه داده همچنین در مقالات [۲، ۵۰ و ۵۳] برای بازشناسی صحنه استفاده شده‌اند.

در شکل ۵-۱ نمونه‌هایی از تصاویر هر دسته و در شکل ۵-۲ میانگین تصاویر را برای هر ۸ دسته را به ترتیب می‌بینیم.

---

۱ coast
۲ forest
۳ highway
۴ insidecity
۵ mountain
۶ opencountry
۷ street
۸ tallbuilding



شکل ۵-۱ نمونه‌هایی از تصاویر دسته‌های مجموعه داده ۸ دسته صحنه خارجی. الف تا ج- ساحل، د تا و- جنگل، ز تا ط- بزرگراه، ی تا ل- درون شهر، م تا س- کوه، ع تا ق- منظره باز، ر تا ت- خیابان، ث تا ذ- ساختمان بلند.

شکل ۵-۲ میانگین تصاویر برای دسته‌های مجموعه داده ۸ دسته صحنه خارجی. به ترتیب از چپ به راست: ساحل، جنگل، بزرگراه، درون شهر، کوه، منظره باز، خیابان، ساختمان بلند. (این تصاویر از میانگین‌گیری بین صدها تصویر از هر دسته به‌دست آمده‌اند) [۵۲].

در این پایان‌نامه از فایل‌های تصویر به همراه فایل‌های تفسیر[۱] آن‌ها استفاده شده است. ابتدا تمام این فایل‌های مجموعه داده، از وب‌گاه لیبل‌می[۲] دانلود و ذخیره شدند. سپس با استفاده از جعبه ابزار لیبل‌می، فایل‌های تصویر و تفسیر به صورت ساختاری به‌ترتیب در آدرس‌های HOMEIMAGES و HOMEANNOTATIONS در برنامه متلب ذخیره می‌شوند. سپس با کمک تابع LMdatabase تصاویر و تفسیرها را در مسیر قرار می‌دهیم. دقت داشته باشید که هم در پوشه تصاویر و هم در پوشه تفسیرها باید تصاویر هر دسته در یک پوشه جداگانه ذخیره شوند.

۵-۳- معیارهای ارزیابی

در مسائل مختلف معیارهای ارزیابی متفاوتی استفاده می‌شود، اما برخی معیارهای ارزیابی به دلیل استاندارد بودن در بسیاری پژوهش‌ها مورد استفاده قرار می‌گیرند. در این پایان‌نامه نیز سعی شده است تا الگوریتم پیشنهادی با برخی از این معیارها مورد ارزیابی قرار گیرد. این معیارهای ارزیابی به شرح زیر می‌باشند:

۱. صحت[۳]: نسبت کل پیش‌بینی‌های درست

---

[۱] annotation
[۲] LableMe
[۳] accuracy

۲. نرخ مثبت درست[۱] (فراخوانی): نسبت تعداد صحنه‌هایی که به درستی در دسته‌ی صحیح خود دسته‌بندی شده‌اند.

۳. مثبت غلط[۲]: نسبت صحنه‌هایی که به طور صحیح دسته‌بندی نشده‌اند و در دسته‌ی دیگری قرار گرفته‌اند.

۴. منفی غلط[۳]: نسبت صحنه‌هایی که جز دسته‌ی موردنظر نیستند، ولی در این دسته قرار گرفته‌اند.

۵. منفی درست[۴]: نسبت صحنه‌هایی که جز دسته‌ی موردنظر نیستند، و به درستی در این دسته قرار نگرفته‌اند.

۶. دقت[۵]: نسبت صحنه‌های درست پیشنهادی که صحیح هستند.

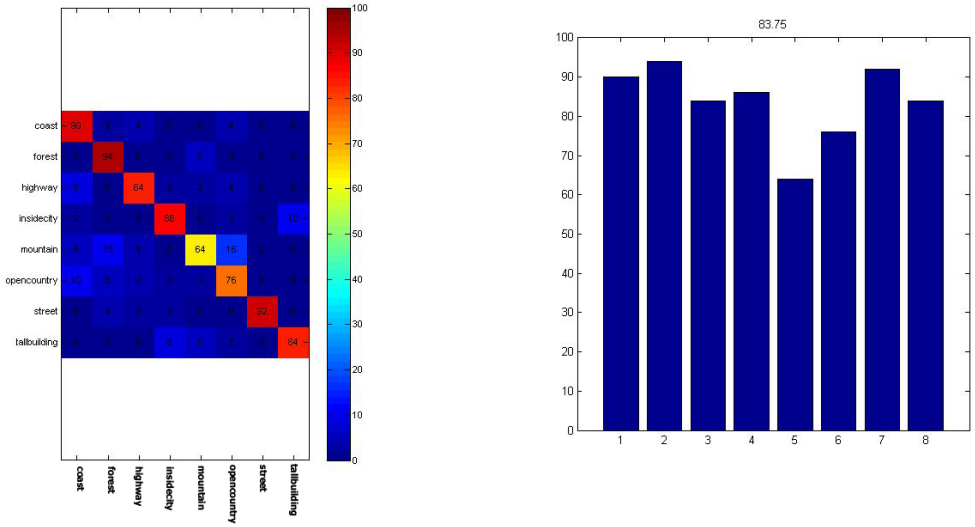
۷. ضریب f[۶]: نسبت حاصلضرب فراخوانی و دقت به مجموع آن‌ها که با استفاده از معادله ۵-۱ زیر به‌دست می‌آید:

$$\text{ضریب } f = (\text{دقت} * \text{فراخوانی})/(\text{دقت} + \text{فراخوانی}) \qquad ۵\text{-}۱$$

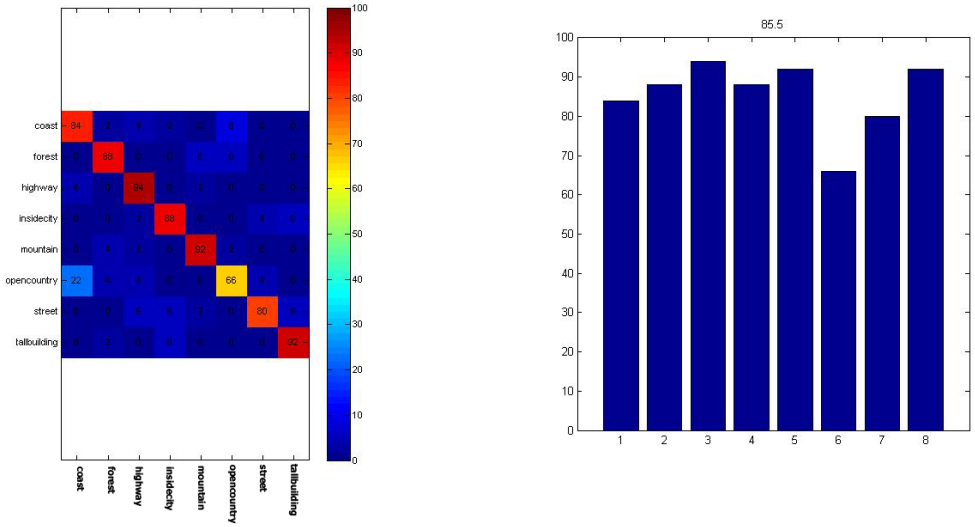
در شکل‌های ۵-۳، ۵-۴ و ۵-۵ به ترتیب ماتریس تقابل نتایج دسته‌بندی الگوریتم‌های جیست نرمال شده، سیفت و الگوریتم پیشنهادی بر روی مجموعه داده ۸ دسته صحنه خارجی نشان داده شده است. در برخی مواقع به دلیل شباهت ظاهری محیط و یا اشیا موجود در صحنه، تصاویر دسته‌های مختلف در عمل دسته‌بندی، با یکدیگر اشتباه گرفته می‌شوند.

---

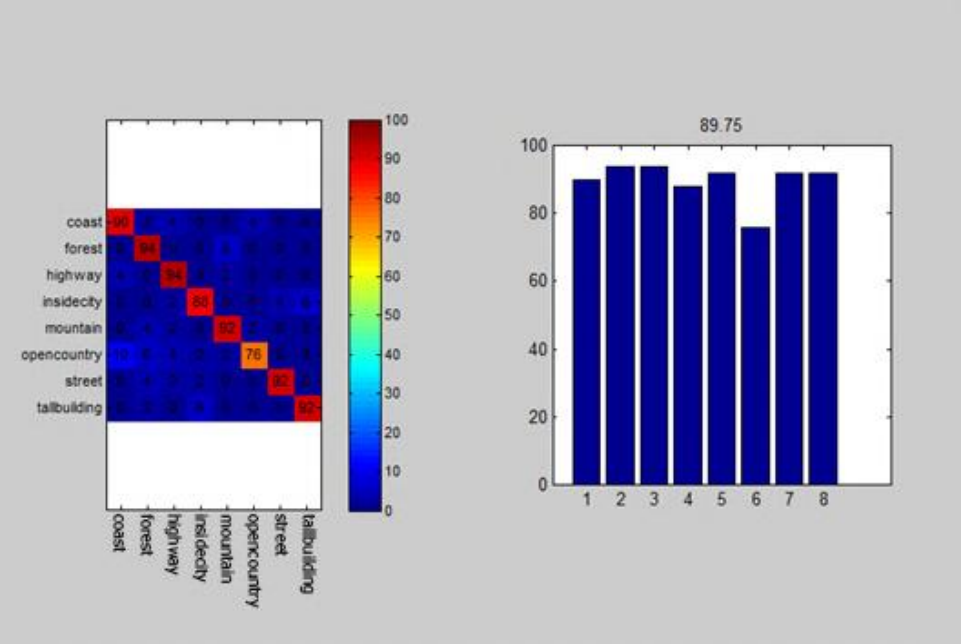
۱ true positive
۲ false positive
۳ false negative
۴ true negative
۵ precision
۶ f-measure



شکل ۵-۳ ماتریس تقابل بین دسته‌های مجموعه داده ۸ دسته صحنه خارجی براساس روش جیست نرمال شده.



شکل ۵-۴ ماتریس تقابل بین دسته‌های مجموعه داده ۸ دسته صحنه خارجی براساس روش سیفت.



شکل ۵-۵ ماتریس تقابل بین دسته‌های مجموعه داده ۸ دسته صحنه خارجی براساس الگوریتم پیشنهادی.

در محور عمودی نام دسته‌ها و در محور افقی هم این نام‌ها برای بازنمایی میزان انطباق بین هر دو دسته قرار گرفته‌اند. هر سلول تعداد صحنه‌هایی که در دسته منطبق افقی قرار دارد و با کلاس‌بند در دسته منطبق عمودی قرار گرفته‌شده را نشان می‌دهد. طیف رنگی کنار نمودار میزان دقت را نشان می‌دهد. و در سمت راست هم نمودار میله‌ای دقت برای هر دسته نشان داده شده است.

باتوجه به ماتریس تقابل می‌توان کلیه معیارهای استاندارد موردنیاز را به دست آورد که نتایج به ترتیب در جداول ۵-۱، ۵-۲ و ۵-۳ برای الگوریتم‌های جیست نرمال شده، سیفت و روش پیشنهادی آورده شده است.

جدول ۵-۱ معیارهای استاندارد برای مجموعه داده ۸ دسته صحنه خارجی با استفاده از الگوریتم جیست نرمال شده (کلیه مقادیر به درصد هستند).

الگوریتم جیست نرمال شده	صحت	مثبت درست (فراخوانی)	منفی درست	مثبت غلط	منفی غلط	دقت	ضریب f
ساحل	۹۰/۰۰	۹۰/۰۰	۸۳/۰۰	۱۵/۷۱	۱۰/۰۰	۸۷/۹۵	۴۲/۰۶
جنگل	۹۴/۰۰	۹۴/۰۰	۸۲/۰۰	۱۷/۴۳	۶/۰۰	۸۲/۴۶	۴۳/۹۳
بزرگراه	۸۴/۰۰	۸۴/۰۰	۸۴/۰۰	۱۶/۰۰	۱۶/۰۰	۸۵/۷۱	۴۲/۴۲
درون شهر	۸۶/۰۰	۸۸/۰۰	۸۳/۰۰	۱۶/۲۹	۱۲/۰۰	۸۶/۰۰	۴۳/۴۹
کوه	۶۴/۰۰	۶۴/۰۰	۸۷/۰۰	۱۳/۱۴	۳۶/۰۰	۸۰/۰۰	۳۵/۵۶
منظره باز	۷۶/۰۰	۷۶/۰۰	۸۵/۰۰	۱۳/۸۶	۲۴/۰۰	۷۳/۰۸	۳۷/۳۶
خیابان	۹۲/۰۰	۹۲/۰	۸۵/۵۷	۱۷/۱۴	۸/۰۰	۱۰۰/۰۰	۴۷/۹۲
ساختمان بلند	۸۴/۰۰	۸۴/۰۰	۸۳/۴۳	۱۶/۰۰	۱۶/۰۰	۸۹/۳۶	۴۳/۳۰

جدول ۵-۲ معیارهای استاندارد برای مجموعه داده ۸ دسته صحنه خارجی با استفاده از الگوریتم سیفت.

ضریب f	دقت	منفی غلط	مثبت غلط	منفی درست	مثبت درست (فراخوانی)	صحت	الگوریتم سیفت
۴۰/۰۰	۷۶/۳۶	۱۶/۰۰	۱۳/۴۳	۸۵/۷۱	۸۴/۰۰	۸۴/۰۰	ساحل
۴۴/۰۰	۸۸/۰۰	۱۲/۰۰	۱۴/۰۰	۸۵/۱۴	۸۸/۰۰	۸۸/۰۰	جنگل
۴۴/۳۴	۸۳/۹۳	۶/۰۰	۱۴/۸۶	۸۴/۲۹	۹۴/۰۰	۹۴/۰۰	بزرگراه
۴۴/۰۰	۸۸/۰۰	۱۲/۰۰	۱۴/۰۰	۸۵/۱۴	۸۴/۰۰	۸۸/۰۰	درون شهر
۴۵/۵۵	۹۰/۲۰	۸/۰	۱۴/۵۷	۸۴/۵۷	۹۲/۰۰	۹۲/۰۰	کوه
۳۶/۲۶	۸۰/۴۹	۳۴/۰۰	۱۰/۸۶	۸۸/۲۹	۶۶/۰۰	۶۶/۰۰	منظره باز
۴۲/۵۵	۹۰/۹۰	۱۰/۰۰	۱۳/۷۱	۸۶/۳۹	۸۸/۰۰	۸۰/۰۰	خیابان
۴۲/۹۹	۸۰/۷۰	۸/۰۰	۱۴/۵۷	۸۴/۵۷	۹۲/۰۰	۹۲/۰۰	ساختمان بلند

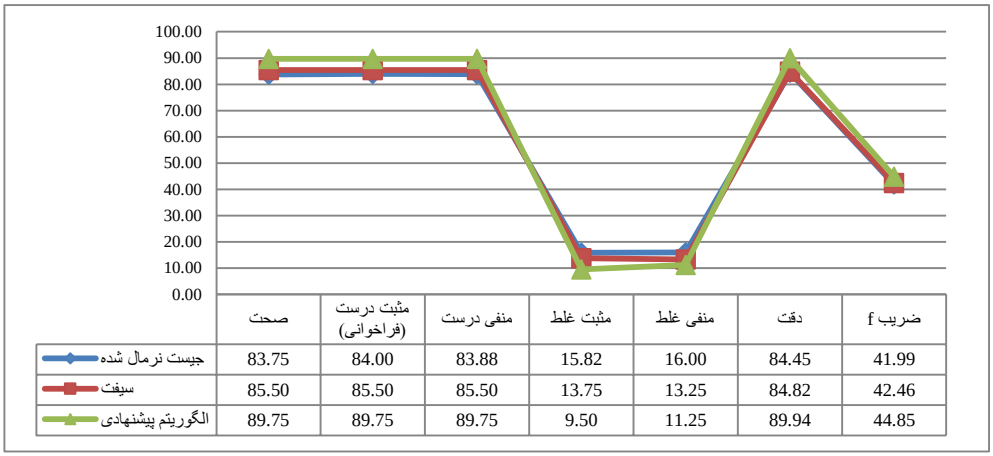
جدول ۵-۳ معیارهای استاندارد برای مجموعه داده ۸ دسته صحنه خارجی با استفاده از الگوریتم پیشنهادی.

ضریب f	دقت	منفی غلط	مثبت غلط	منفی درست	مثبت درست (فراخوانی)	صحت	الگوریتم پیشنهادی
۴۴/۱۲	۸۶/۵۴	۱۸/۰۰	۱۰/۲۹	۸۹/۷۱	۹۰/۰۰	۹۰/۰۰	ساحل
۴۴/۳۴	۸۳/۹۳	۶/۰۰	۱۰/۸۶	۸۹/۱۴	۹۴/۰۰	۹۴/۰۰	جنگل
۴۵/۱۹	۸۷/۰۴	۶/۰۰	۱۰/۸۶	۸۹/۱۴	۹۴/۰۰	۹۴/۰۰	بزرگراه
۴۴/۴۵	۸۹/۸۰	۱۲/۰۰	۱۰/۰۰	۹۰/۰۰	۸۸/۰۰	۸۸/۰۰	درون شهر
۴۶/۰۰	۹۲/۰۰	۸/۰۰	۸/۸۶	۸۹/۴۳	۹۲/۰۰	۹۲/۰۰	کوه
۴۱/۳۱	۹۰/۴۸	۲۴/۰۰	۷/۴۳	۹۱/۷۱	۷۶/۰۰	۷۶/۰۰	منظره باز
۴۶/۹۴	۹۵/۸۳	۸/۰۰	۸/۸۶	۸۹/۴۳	۹۲/۰۰	۹۲/۰۰	خیابان
۴۶/۴۷	۹۳/۸۸	۸/۰۰	۸/۸۶	۸۹/۴۳	۹۲/۰۰	۹۲/۰۰	ساختمان بلند

۵-۴- ارزیابی نتایج

در شکل ۵-۶ مقایسه بین روش‌های جیست نرمال شده، سیفت و الگوریتم پیشنهادی براساس کلیه معیارهای استاندارد نامبرده انجام شده است. همان‌طور که مشاهده می‌شود روش پیشنهادی نسبت به هردو روش عملکرد بهتری دارد.

برای ارزیابی این الگوریتم از چند الگوریتمی که عملکرد خوبی بر روی مجموعه داده موردنظر داشته‌اند، استفاده می‌کنیم. ابتدا نتایج روش جیست که روش پایه است و سپس دیگر روش‌ها که اغلب از ترکیب دو ویژگی به‌دست آمده‌اند، مشاهده می‌شود.



	صحت	مثبت درست (فراخوانی)	منفی درست	مثبت غلط	منفی غلط	دقت	ضریب f
جیست نرمال شده	83.75	84.00	83.88	15.82	16.00	84.45	41.99
سیفت	85.50	85.50	85.50	13.75	13.25	84.82	42.46
الگوریتم پیشنهادی	89.75	89.75	89.75	9.50	11.25	89.94	44.85

شکل ۵-۶ مقایسه عملکرد روش‌های جیست نرمال شده، سیفت و الگوریتم پیشنهادی براساس معیارهای استاندارد.

جدول ۵-۴ نتایج عملی چند روش روی مجموعه داده ۸ دسته صحنه خارجی.

الگوریتم	صحت (%)
جیست نرمال شده [۲]	۸۳/۷۵
جیست نوین[۱] [۳۹]	۸۶/۶۰
سنتریست[۲] [۵۴]	۸۶/۲۰
جیست سی ام سی تی[۳] [۵۵]	۸۵/۸۲
ام سی تی مکانی[۴] [۵۰]	۸۷/۶۵
سیفت [۴۰]	۸۵/۵۰
جیست سی ام سی تی و ام سی تی مکانی [۵۰]	۸۸/۹۵
روش پیشنهادی (جیست نرمال شده و سیفت)	**۸۹/۷۵**

طبق جدول بالا بین روش‌های نامبرده بیشترین دقت موردنظر مربوط به آخرین روش ارائه شده یعنی روش ترکیبی برمبنای ویژگی‌های جیست سی ام سی تی و ام سی تی مکانی می‌باشد؛ اما، بردارهای ویژگی به کار گرفته در این روش طبق مقاله [۵۳] ابعاد بزرگی دارند. بردار ویژگی جیست سی ام سی تی از دو بردار ویژگی جیست و سی ام سی تی (CMCT)[۵] تشکیل شده است.

---

[۱] Novel GIST
[۲] Centrist
[۳] GISTCMCT
[۴] Spatial MCT
[۵] contextual mean census transform

همان‌طور که در فصل ۴ گفته شد، ویژگی جیست یک بردار ۵۱۲ بعدی است. ویژگی سی ام سی تی مشخصه‌های ساختاری را از توزیع ساختارهای محلی به‌دست می‌آورد که به کلاس‌بندی محیط‌های مصنوعی، ازجمله، محیط‌های داخلی کمک می‌کند. این بردار هم ۵۱۲ بعدی است. بردار ویژگی جیست سی ام سی تی از این دو بردار یک بردار جدید ایجاد می‌کند. پس، بردار حاصل ۱۰۲۴ بعدی است. بردارهای ام سی تی که شامل مقادیر میانگین تبدیل سنسوس[۱] در همه بلوک‌های به‌دست آمده است، سپس به هم متصل می‌شوند تا یک بردار ویژگی نهایی را تشکیل دهند. بعد از به‌دست آوردن بردار نهایی، از تحلیل مولفه اساسی[۲] (PCA) برای کاهش ابعاد بردار ویژگی استفاده می‌شود. برای تشکیل این بردار ۳۱ بلوک تولید می‌شود که ابعاد هریک از ۵۱۲ به ۴۰ کاهش یافته است. بردار نهایی ۳۱×۴۰ یعنی ۱۲۴۰ بعد دارد[۵۳].

بزرگ بودن ابعاد بردار ویژگی موجب اشغال فضای زیادی از حافظه و همچنین افزایش زمان محاسبه این بردارها برای تصاویر مجموعه داده می‌شود. برای حل این مسئله، در این پایان‌نامه راه‌حلی بر مبنای ویژگی‌های جیست و سیفت ارائه شد؛ که همان‌طور که در فصل ۴ اشاره شد علاوه بر کاهش اندازه بردار ویژگی و درنتیجه کاهش زمان اجرای الگوریتم، میزان دقت کلاس‌بندی برای مجموعه داده موردنظر با توجه به آخرین نتایج به‌دست آمده افزایش داده است.

۵-۵- جمع‌بندی

در این فصل نتایج تجربی اجرای الگوریتم پیشنهادی ارائه شد و با چند الگوریتم مشهور در این زمینه مقایسه گردید. ابتدا مجموعه داده‌ی استفاده شده در روش پیشنهادی معرفی شده و در ادامه نتایج دسته-بندی بر روی این مجموعه داده با استفاده از روش پیشنهادی ارائه شد و با الگوریتم‌های مشابه مقایسه شد.

[۱] mean census transform
[۲] principal component analysis

فصل ششم: نتیجه‌گیری و پیشنهادها

در این پایان‌نامه ابتدا به معرفی یک سیستم ادراک صحنه و سپس به معرفی روش‌های مختلف بازشناسی صحنه پرداخته شد. بعضی از این روش‌ها از ویژگی‌های محلی و بعضی از ویژگی‌های سراسری تصویر استفاده می‌کنند. هریک از این روش‌ها، مزایا و معایب خود را دارند. روش‌های نوین نیز از ترکیب این دو روش استفاده می کنند.

روشی که در این پایان‌نامه به آن پرداخته شد روشی براساس ترکیب ویژگی‌های سراسری و محلی تصویر است. با توجه به این‌که ویژگی‌های سراسری کل صحنه را به عنوان یک موجودیت درنظر می‌گیرند و برخی دسته صحنه‌ها در عین این‌که متفاوتند اما ساختار کلی یکسان و در نتیجه ویژگی‌های سراسری مشابهی دارند، موجب می‌شوند بازشناسی با ویژگی‌های سراسری به سادگی برای آن‌ها انجام نشود و نتایج مطلوبی حاصل نشود. مثلاً دو دسته صحنه ساحل و منظره باز هردو صحنه‌های بازی هستند که در نیمه بالایی تصویر معمولاً بدون بافت و در نیمه پایینی دارای بافت هستند. همچنین این روش‌ها برای دسته صحنه‌هایی که ساختار پیچیده دارند همچون ساختمان بلند عملکرد خوبی ندارند. به علاوه، روش‌های مبتنی‌بر ویژگی-های محلی هم به دلیل قطعه‌بندی تصویر و اعمال الگوریتم روی زیرتصویرها برای بازشناسی صحنه‌های که طرح کلی ساده‌ای دارند، عملکرد خوبی ندارد. روش‌های ترکیبی ویژگی‌های سراسری و محلی می‌توانند در این موارد خاص کمک کنند؛ زیرا سعی می‌کنند از ارتباط اجزای تصویر برای بازشناسی استفاده کنند. در قسمت ارزیابی نشان داده شد که نتایج الگوریتم پیشنهادی امیدبخش است و می‌توان از این روش در شرایط مختلف استفاده کرد.

در آینده به استخراج ویژگی‌های سراسری و محلی دیگر و ترکیب آن‌ها با هم برای تشخیص بهتر صحنه خواهیم پرداخت. همچنین از آنجایی که می‌دانیم ویژگی‌های سراسری و بخصوص جیست برای دسته صحنه‌های داخلی که پیچیدگی بیشتری نسبت به دسته‌های خارجی دارند، به خوبی جواب نمی‌دهند؛ سعی می‌کنیم نتایج را برای این مجموعه داده‌ها و همچنین مجموعه داده‌های بسیار بزرگ‌تر مانند مجموعه داده SUN بهبود دهیم.

مراجع

[۱] Computational Perception & Cognition. [Online].
http://cvcl.mit.edu/scene_understanding.html

[۲] A. Oliva, A. Torrolba, "Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope, "International Journal of Computer Vision", Vol. 42, No. 3, pp. 145-175, 2001.

[۴] Y. Zhao, S. Zhu, "Scene Parsing By Integrating Function, Geometry and Appearance Models", CVPR, pp. 4321- 4328, 2013.

[۵] R. Salakhutdinov, A. Torralba, J. Tenenbaum, "Learning to Share Visual Appearance for Multiclass Object Detection", IEEE Conference on CVPR, pp. 1481 – 1488, 2011, Providence, RI.

[۶] S. Jialin Pan, Q. Yang, "A survey on transfer learning", IEEE Transactions on Knowledge and Data Engineering, Vol. 22, No. 10, pp. 1345 - 1359, 2010.

[۷] J. J. Lim, R. Salakhutdinov, and A. Torralba, "Transfer Learning by Borrowing Examples for Multiclass Object Detection", NIPS, 2011, Granada, Spain.

[۸] K. Grauman and B. Leibe, "Visual Object Recognition: Synthesis Lectures On Computer Vision", Morgan and Claypool Publishers, 2011.

[۹] KAMADA T., KAWAI S., "A Simple Method for Computing General Position in Displaying Three-dimensional Objects", Computer Vision, Graphics and Image Processing, Vol. 41, 1988.

[۱۰] A. Oliva, "Gist of the scene", Encyclopedia of Neurobiology of Attention. L. Itti, G. Rees, and J.K. Tsotsos (Eds.), Elsevier, San Diego, CA, pp. 251-256, 2005.

[۱۱] F. Bermond, "Scene Understanding: Perception, Multi- sensor Fusion, Spatio- temporal Resoning, And Activity Recognition", HDR University de Nice- Sophia Antipolis, July 2007.

[۱۲] S. Wang, Y.Wang, SC. Zhu, "Hierarchical Space Tiling For Scene Modeling", Computer Vision, ACCV, Vol. 7725, pp. 796- 810, 2013.

[۱۳] J. Xiao, J.Hays, B. C. Russel, G. Patterson, K. A.Ehinger, A. Torrolba, A. Oliva, "Basic Level Scene Understanding; Categories, Attributes and Structures", Frontiers in Perception Science, , Vol. 4, No. 506, 2013.

[۱۴] M. R. Greene, A. Oliva, "Recogniton of Natural Scenes From Global Properties: Seeing The Forest Without Representing The Trees", Cognitive Psychology, No. 58, pp. 137- 176, 2009.

[۱۵] L. Li, H. Su, E. P. Xing, L. Fei-Fei, "Object Bank: A High-Level Image Representation for Scene Classification and Semantic Feature Sparsification", Proceeding of Natural Information Processing Systems (NIPS), pp. 1378-1386, 2010.

[۱۶] J. Xiao, H. Hays, A. Ehinger, A. Oliva, A. Torrlba, "SUN Database: Large-scale Scene Recognition From Abbey to Zoo", Proceedings of the 23rd IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3485- 3492, IEEE Computer Society, 2010.

[۱۷] K. Murphy, A. Torrolba, D. Eaton, W. Freeman, "Object Detection and Localization Using Local and Global Features", Toward Category Level Object Recognition, Lecture Note in Computer Science, Vol. 4170, pp. 382- 400. 2006.

[۱۸] A. Oliva, "Scene Perception", Chapter (51) in the New Visual Neurosciences. Eds John S. Werner and Leo. M. Chalupa, pp. 725- 732, 2013.

[۱۹] Y. Liu, D.S. Zhang, G. Lu, W.-Y. Ma, "Region-based image retrieval with perceptual colors", In Proc. of the Pacific-Rim Multimedia Conference (PCM), pp. 931-938, 2004.

[۲۰] A. Bosch, A. Zisserman, X. Munoz, "Scene classification via pLSA", In Proc. of the European Conference on Computer Vision, vol. 4, pp. 517-530, 2006.

[۲۱] P. Quelhas, F. Monay, J. Odobez, D. Gatica-Perez, T. Tuytelaars, L. Van Gool, "Modeling scenes with local descriptors and latent aspects", In Proc. of the IEEE International Conference on Computer Vision, Vol. 1, pp. 883-890, 2005.

[۲۲] J. Luo, A. Savakis, "Indoor vs outdoor classification of consumer photographs using low-level and semantic features", International Conference on Image Processing (ICIP), vol. 2, pp. 745-748, 2001.

[۲۳] R. Fergus, P. Perona, and A. Zisserman, "Object class recognition by unsupervised scale-invariant learning", IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 264-271, 2003.

[۲۴] C. Carson, M. Thomas, S. Belongie, J. Hellerstein and J. Malik, "Blobworld: A system for region-based image indexing and retrieval", In Third Int. Conf. on Visual Information Systems, Springer-Verlag, 1999.

[۲۵] M. Gorkani and R. Picard, "Texture orientation for sorting photos at a glance", in Proc. of the IEEE Conference on Pattern Recognition (ICPR), vol. 1, pp. 459-464, October 1994.

[۲۶] L. Fei-Fei, and P. Perona, "A Bayesian Hierarchical model for learning natural scene categories", IEEE Proceedings in Computer Vision and Pattern Recognition, vol. 2, pp. 524-531, 2005.

[۲۷] S. Lazebnik, C. Schmid, J. Ponce, “Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories”, In Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2169-2178, 2006.

[۲۸] R. Zhang and Z. Zhang, “Hidden semantic concept discovery in region based image retrieval”, In Proc. of the Conference on Computer Vision and Pattern Recognition (CVPR), 2004.

[۲۹] S. Kumar and M. Hebert, “A hierarchical field framework for unified context-based classification”, In Proc. of the ICCV, vol. 2, pp.1284-1291, 2005.

[۳۰] C.P. Town, D. Sinclair, “Content-based image retrieval using semantic visual categories”, Society for Manufacturing Engineers, Technical Report, 2001.

[۳۱] Y. Li, J. Bilmes, L.G. Shapiro, “Object class recognition using images of abstract regions”, In Proc. of the International Conference on Pattern Recognition, 2004.

[۳۲] Y. Li, L.G. Shapiro, J. Bilmes, “A generative/discriminative learning algorithm for image classification”, In Proc. of the International Conference on Computer Vision (CVPR), 2005.

[۳۳] L. W. Renninger and J. Malik, “When is scene identification just texture recognition?”, Vision Research, vol. 44, pp. 2301-2311, 2004.

[۳۴] S. Tong, E. Chang, “Support vector machine active learning for image retrieval”, In Proc. of the ACM International Conference on Multimedia, pp. 107-118, 2001.

[۳۵] A. Vailaya, A. Jain, and H. Zhang, “On image classification: City vs. landscape”, Pattern Recognition, vol. 31(12), pp. 1921-1935, 1998.

[۳۶] D. Hoiem, A. A. Efros, and M. Hebert, “Closing the loop on scene interpretation”, In Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2008.

[۳۷] A. Oliva and A. Torralba, “Scene-centered description from spatial envelope properties”, In Proc. of the 2nd international workshop on biologically motivated computer vision, pp. 263-272, 2002.

[۳۸] D. G. Lowe, "Object recognition from local scale-invariant features," in Computer vision, 1999. The proceedings of the seventh IEEE international conference on, 1999, pp. 1150-1157.

[۳۹] Julia Vogel and Bernt Schiele, “A semantic typicality measure for natural scene categorization,” Pattern Recognition Symposium DAGM, 2004.

[۴۰] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," International journal of computer vision, vol. 60, pp. 91-110, 2004.

[۴۱] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, 2005, pp. 886-893.

[۴۲] E. B. Sudderth, A. Torralba, W. T. Freeman, and A. S. Willsky, "Learning hierarchical models of scenes, objects, and parts," in Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on, 2005, pp. 1331-1338.

[۴۳] R. C. Gonzalez, R. E. Woods, and S. L. Eddins, Digital image processing using MATLAB: Pearson Education India, 2004.

[۴۴] A. Oliva and A. Torralba, "Building the gist of a scene: The role of global image features in recognition," Progress in brain research, vol. 155, pp. 23-36, 2006.

[۴۵] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman, "LabelMe: a database and web-based tool for image annotation," International journal of computer vision, vol. 77, pp. 157-173, 2008.

[۴۶] A. Quattoni and A. Torralba, "Recognizing indoor scenes," Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, 2009.

[۴۷] Hamed Kiani and Terence Sim, "Scene Classification: Computational and Cognitive Approaches", Graduate Research Paper (GRP), School of Computing, NUS, 2011.

[۴۸] A. Torralba and A. Oliva, "Statistics of natural image categories", Network: computation in neural systems, Vol. 14, 391-412. 2003. Online at: stacks.iop.org/Network/14/391.

[۴۹] A. Torralba, A. Oliva, "Depth estimation from image structure", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 24(9): 1226-1238. 2003.

[۵۰] de Souza Gazolli, Kelly Assis; Salles, Evandro Ottoni Teattini, "Using holistic features for scene classification by combining classifiers" , Journal of WSCG, vol. 21, no. 1, p. 41-48, 2013.

[۵۱] A. Oliva, A. Torralba, A. Geurin-Dugue, J. Herault, "Global Semantic Classification of Scenes using Power Spectrum Templates", Challenge of Image Retrieval (CIR99), Elect, Work. In Computing Series, Springer-Verlag, Newcastle, 1999.

[۵۲] http://people.csail.mit.edu/torralba/code/spatialenvelope/

[۵۳] X. Meng and Z. Wang, " Rapid scene categorization using novel gist model:, In Information 80Engineering and Computer Science (ICIECS), 2010, 2nd International Conference on, 2010.

[۵۴] J. Wu and J. M. Rehg, "Centrist: A visual descriptor for scene categorization", IEEE Trans.Pattern Anal. Mach. Intell., p.p. 1489–1501, 2011.

[۵۵] X. Meng, Z. Wang and L. Wu, "Building global image features for scene recognition", Journal on Pattern Recognition, vol. 45, p.p. 373-380, 2012.

Abstract

Scene recognition is a challenging issue in computer vision and includes two essential parts: Feature extraction and classification. Performance of the recognition method is highly dependent on feature extraction.

Prior methods can be characterized by the level and the scale of feature extraction, respectively. In level point of view, an image can be represented by low level (e.g. color, texture and edge) or contextual level (e.g. scene's 3D layout). In the case of extraction scale, different methods use different scale of image processing for feature extraction, from local (blocks, objects, regions, blobs, pixels) to global (whole image, holistic) scale. In spite that some proposed frameworks performed scene recognition using low level features, the "semantic gap" between low level features and high level semantic concepts, forces researchers to use contextual level features at higher level of description. Also, Due to low attention of methods based global features to parts of images, using of hybrid methods by combining both global and local features is more efficient.

This paper proposes a novel algorithm on 8- outdoor scene categories data set (that includes 2688 images from coast, forest, highway, insidecity, mountain, open country, street and tall building categories) based on combination classifiers. First, we extract two vectors based on gist and sift features for each scene image and classify all images to categories separately with a support vector machine classifier. Then combine the results to obtain final class. Experimental results on 8-outdoor scene categories dataset shows that the proposed method improve accuracy and computational time against previous methods without increasing the final size of the feature vector.

Experimental results on this dataset show that the proposed method outperforms similar methods in terms of accuracy by taking advantage from the qualities of the two descriptors without increasing size of feature vectors.

Keywords gist, outdoor scene, scene recognition, sift, support vector machine

Kharazmi University of Tehran

Faculty of Engineering

M. Sc. Thesis

Computer Engineering- Artificial Intelligence

Scene recognition using hybrid global and local image features

By:

Fatemeh Ghanbari Adivi

Supervisor:

Dr. Jamshid Shanbehzadeh

Advisor:

Dr. Zeinab Ghassabi

2015, January