
بسم الله الرحمن الرحیم

دانشکده مهندسی کامپیوتر

رساله دکتری مهندسی هوش مصنوعی

دسته‌بندی محتوایی تصویر بر پایه مدل‌های موضوعی

نگارنده

علی قنبری سرخی

استاد راهنما

دکتر حمید حسن‌پور

استاد مشاور

دکتر منصور فاتح

بهمن ۱۳۹۷

تقدیم به پدر گرانقدر و مادر مهربانم

آن فرشته‌هایی که از خواسته‌هایشان گذشتند، سختی‌ها را به جان خریدند و خود را سد مشکلات و ناملایمات زندگی کردند تا رسیدن به جایگاهی که اکنون در آن می باشم هموار شود. همراهان عزیزی که نه می‌توانم موهای سپیدشان را به رنگ زمان جوانی کنم و نه برای ترمیم دست‌های پینه بسته‌یشان مرهمی گذارم. باری‌تعالی به افکارم بیاموز تا زمانی که توان همنشینی با این موهبت‌های الهی برای من مقدور است عصای دست هایشان بوده و لحظه لحظه شکرگزار تلاش های بی وقفه و بی مانندشان باشم....

تقدیم به همسر نازنینم

همان همدمی که با وفاداری و رفاقت تمام معنای خود آرام کننده‌ی روح و روانم بود و مانند کوه در کنار من بر قامت مشکلاتم ایستاد تا این مسیر طولانی وسخت، کوتاه و شیرین به نظر آید.

شاید تقدیر کمترین واژه برای بیان احساسات ناگفتنی از این نازنین بی‌مانند باشد که در این تلاطم‌های زندگی از خواسته‌های خود کاست و بر داشته‌های من افزود.

معنای دوست داشتن و رفاقت در قلب مهربان این ستاره کهکشانم تجلی می‌یابد؛ باشد که این همراهی تا نفس‌های آخرمان پیوندی ناگسستنی داشته باشد.

تقدیر و تشکر

سپاس خدای را که سخنوران، در ستودن او بمانند و شمارندگان، شمردن نعمت‌های او ندانند و کوشندگان، حق او را گزاردن نتوانند.

بر خود لازم می‌دانم که از کلیه سرورانی که بنده را در این پژوهش یاری نمودند، تشکر و قدرانی نمایم. بدون شک جایگاه و منزلت معلم، اجّل از آن است که در مقام قدردانی از وی بتوان حرفی زد.

در ابتدا بایستی از آقای دکتر **حمید حسن‌پور** که زحمت راهنمایی این رساله را برعهده گرفتند، تشکر و قدردانی نمایم. ایشان با صبر و حوصله و همچنین راهنمایی‌های دقیق و ارزشمند، مرا در این پژوهش یاری کردند. بدون شک بدون راهنمایی‌های ایشان رسیدن به این مقصود ممکن نبود.

همچنین از استاد فرزانه، آقای دکتر منصور فاتح که زحمت مشاوره این رساله را بر عهده گرفتند، کمال تشکر و قدردانی را دارم. نظرات و راهنمایی‌های ارزشمند ایشان روشنگر مسیر من در این پژوهش بوده است.

در پایان از تمامی کسانی که به هر نحو ممکن مرا در این پژوهش یاری کردند، کمال تشکر و سپاسگزاری را دارم.

تعهد نامه

اینجانب **علی قنبری سرخی** دانشجوی دوره دکتری رشته مهندسی کامپیوتر - هوش مصنوعی دانشکده مهندسی کامپیوتر دانشگاه صنعتی شاهرود نویسنده پایان‌نامه **دسته‌بندی محتوایی تصویر بر پایه مدل‌های موضوعی** تحت راهنمائی دکتر حمید حسن‌پور متعهد می شوم.

- تحقیقات در این پایان‌نامه توسط اینجانب انجام شده است و از صحت و اصالت برخوردار است.
- در استفاده از نتایج پژوهشهای محققان دیگر به مرجع مورد استفاده استناد شده است.
- مطالب مندرج در پایان‌نامه تاکنون توسط خود یا فرد دیگری برای دریافت هیچ نوع مدرک یا امتیازی در هیچ جا ارائه نشده است.
- کلیه حقوق معنوی این اثر متعلق به دانشگاه صنعتی شاهرود می‌باشد و مقالات مستخرج با نام « دانشگاه صنعتی شاهرود » و یا « Shahrood University of Technology» به چاپ خواهد رسید.
- حقوق معنوی تمام افرادی که در به‌دست آمدن نتایج اصلی پایان‌نامه تأثیرگذار بوده اند در مقالات مستخرج از پایان‌نامه رعایت می گردد.
- در کلیه مراحل انجام این پایان‌نامه ، در مواردی که از موجود زنده (یا بافتهای آنها) استفاده شده است ضوابط و اصول اخلاقی رعایت شده است.
- در کلیه مراحل انجام این پایان‌نامه، در مواردی که به حوزه اطلاعات شخصی افراد دسترسی یافته یا استفاده شده است اصل رازداری ، ضوابط و اصول اخلاق انسانی رعایت شده است.

تاریخ

امضای دانشجو

مالکیت نتایج و حق نشر

- *کلیه حقوق معنوی این اثر و محصولات آن (مقالات مستخرج، کتاب، برنامه های رایانه ای، نرم افزار ها و تجهیزات ساخته شده است) متعلق به دانشگاه صنعتی شاهرود می‌باشد. این مطلب باید به نحو مقتضی در تولیدات علمی مربوطه ذکر شود.*
- *استفاده از اطلاعات و نتایج موجود در پایان‌نامه بدون ذکر مرجع مجاز نمی‌باشد.*

چکیده

با افزایش روز افزون تصاویر دیجیتال و تبادل آن‌ها در شبکه اینترنت، مسئله طبقه‌بندی تصاویر به یکی از نیازهای اساسی دنیای دیجیتال تبدیل شده است. توصیف خودکار محتوا یکی از مشکلات اساسی دسته‌بندی تصاویر است که این امر سبب ارتباط بیشتر بین بینایی ماشین و پردازش تصویر شده است. الگوریتم‌های توصیف خودکار محتوا کاربردهای فراوانی در سیستم‌های تحت وب و موتورهای جستجو نظیر فیلتر نمودن تصاویر نامتعارف، تشخیص موضوعی تصویر و تشخیص رفتار انسانی دارند.

مسئله توصیف خودکار محتوا را می‌توان در دو دیدگاه مطالعه کرد. در دیدگاه اول مجموعه داده تصاویر، مجموعه‌ای کوچک با کاربرد منحصر به فرد مانند شناسایی تصاویر نامتعارف است. شناسایی تصاویر نامتعارف، یک مسئله توصیف با دو کلاس و پیچیدگی پایین است. در این رساله، معماری جدیدی برای شبکه عصبی عمیق به منظور تشخیص تصاویر نامتعارف پیشنهاد می‌شود. تاکید معماری پیشنهادی استخراج ویژگی‌های سطح بالا از بدن انسان در تصاویر نامتعارف است. نتایج آزمایش روش پیشنهادی بر روی دو مجموعه داده نشان دهنده بهبود عملکرد تا حدود ۴٪ نسبت به روش‌های مطرح شده در سال‌های اخیر است.

در دیدگاه دوم، مجموعه داده‌ها شامل تصاویری با تعداد کلاس‌های بیشتری از صحنه‌های متفاوت است. در این رساله، روشی به منظور توصیف خودکار تصاویر بر اساس استخراج نواحی معنایی سطح بالا و استفاده از این نواحی برای استخراج برچسب‌های اشیاء پیشنهاد می‌شود. روش ارائه شده شامل چندین مرحله می‌باشد. درمرحله نخست نواحی و برچسب اشیاء شناسایی می‌شود. در این راستا، یک معماری جدید بر پایه شبکه عمیق بهبود یافته از روش R-FCN ارائه می‌شود. در این روش از یک تابع زیان جدید که تاکنون برای شناسایی اشیاء استفاده نشده، استفاده می‌شود. با توجه به روش ارائه شده در این مرحله، زمان آموزش و آزمایش بهبود یافته است. یک گام مهم در توصیف خودکار تصاویر، استخراج موضوعات نهفته در صحنه تصویر می‌باشد. بنابراین در مرحله بعدی، موضوعات نهفته برپایه نواحی معنایی سطح بالای استخراج شده به منظور استخراج مدل موضوعی به کمک روش‌های مجموعه کلمات و k-means بهبود یافته استخراج می‌شود. در ادامه با توجه به مجموعه کلمات بصری استخراج شده از نواحی کاندیدا، محل قرار گرفتن آنها و برچسب اشیاء از روش عمیق تخمین‌گر توزیع‌شده خودرگرسیو عصبی اسناد برای محاسبه موضوعات نهفته در هر تصویر استفاده می‌شود. در نهایت با ترکیبی از این ویژگی‌های سطح بالا که نشان دهنده محتوای بصری تصویر می‌باشند دسته‌بندی صحنه تصویر انجام می‌گردد. روش پیشنهادی از لحاظ زمان و صحت بر روی مجموعه داده‌های Scene15، UIUC Sports و MIT-67 به ترتیب با ۱۵، ۸ و ۶۷ دسته متفاوت مورد ارزیابی قرار گرفت. نتایج نشان دهنده بهبود عملکرد روش پیشنهادی از لحاظ صحت و زمان مرحله آزمون در این مجموعه داده‌ها می‌باشد.

کلمات کلیدی: دسته‌بندی صحنه تصویر، شبکه‌عصبی عمیق، استخراج محتوا، مدل‌سازی موضوعی

لیست مقالات مستخرج از رساله

مجلات عملی – پژوهشی

[مقاله ۱]. علی. قنبری سرخی ، منصور. فاتح و حمید. حسن‌پور، « **تشخیص و فیلترینگ هوشمند تصاویر نامتعارف به کمک شبکه‌های عصبی عمیق** »، فصل‌نامه علمی-پژوهشی پردازش علائم و داده‌ها، جلد ۱۵، شماره ۲، صفحات ۵۵-۶۸، ۱۳۹۷.

[مقاله ۲]. علی. قنبری سرخی ، حمید. حسن‌پور و منصور. فاتح، « **بهبود شبکه عمیق** R-FCN **در آشکارسازی و برچسب‌زنی اشیاء**»، مجله ماشین بینایی و پردازش تصویر. (پذیرش)

[Paper 3]. A. Ghanbari sorkhi, H. Hassanpour and M. Fateh., " A Comprehensive System for Image Scene Classification", Multimedia Tools and Applications. (Under review)

مجلات عملی – ترویجی

[مقاله ۴]. علی. قنبری سرخی ، حمید. حسن‌پور و منصور. فاتح، « **انتخاب ناحیه‌های کاندیدا در سیستم‌های تشخیص و شناسایی اشیاء**»، مجله محاسبات نرم، جلد ۵، شماره ۲، صفحات ۳۴-۴۷، ۱۳۹۵.

کنفرانس

[مقاله ۵]. علی. قنبری سرخی ، حمید. حسن‌پور و منصور. فاتح، « **دسته‌بندی صحنه تصویر بر پایه روش ناحیه کاندیدا و شبکه‌های** Boosting »، اولین *کنگره و نمایشگاه بین‌المللی علوم و تکنولوژی‌های نوین*، بابل، ایران، ۱۳۹۷.

فهرست مطالب

فهرست تصاویر

فهرست جداول

واژه‌های اختصاری

علامت اختصاری	واژه اصلی
SUN	Scene UNderstanding
BoW	Bag-of-Words
SIFT	Scale-Invariant Feature Transform
HoG	Histogram of Gradients
SPM	Spatial Pyramid Matching
KPCA	Kernel Principal Component Analyses
RPN	Region Proposal Network
GPU	Graphics Processing Unit
SP	SuperPixel
CPMC	Constrained Parametric Min-Cuts
MCG	Multiscale Combinatorial Grouping
pLSA	probabilistic Latent Semantic Analysis
LDA	Latent Dirichlet Allocation
DBM	Deep Boltzmann Machine
DocNADE	Document Neural Autoregressive Distribution Estimator
McMC	Markov chain Monte Carlo
ReLU	Rectified Linear Unit
FC	Fully Connected
ILSVRC	ImageNet Large Scale Visual Recognition Competition
SVD	Singular Value Decomposition
mAP	mean Average Precision
SBIR	Semantic Based Image Retrieval

۱ مقدمه‌ای بر دسته‌بندی محتوایی تصویر

۱ - ۱ مقدمه

با رشد تکنولوژی کامپیوتر، اهمیت فوق‌العاده اطلاعات چندرسانه‌ای، وجود آرشیوهای بزرگ دیجیتال و رشد خیلی سریع شبکه گسترده جهانی، در سال‌های اخیر تحقیقات بسیاری به منظور ایجاد ابزارهای مناسب جهت دسته‌بندی تصویر انجام شده است. امروزه مردم حجم انبوهی از داده‌های الکترونیکی را در اختیار دارند که بخش عظیمی از آن‌ها را تصاویر تشکیل می‌دهند. ذخیره این تصاویر منجر به تولید پایگاه‌داده حجیمی از تصاویر می‌شود که جستجو توسط کاربر را در این پایگاه داده سخت خواهد بود. بنابراین نیاز به الگوریتم‌های خودکار و کارآمد برای جستجو در پایگاه‌های تصویری و دسته‌بندی تصاویر دلخواه ضروری به نظر می‌رسد.

هدف از طرح مسئله‌ی دسته‌بندی تصاویر، تشخیص محتوای یک تصویر و تعیین دسته‌ای است که تصویر به آن تعلق دارد. برای آنکه بتوان محتوای یک تصویر را شناسایی کرد، لازم است اطلاعاتی از تصویر استخراج شود و در گام بعدی این اطلاعات مورد پردازش قرار گیرد. برای پردازش این اطلاعات روش‌های مختلفی ارائه شده است که هرکدام ویژگی‌های خاص خود را دارند. منظور از دسته‌بندی تصویر، تشخیص موضوع اصلی تصویر و مفهومی است که به بیننده می‌رساند. در نتیجه می‌توان تصویر را در یکی از دسته‌هایی قرار داد که از پیش تعریف شده است. یکی از کاربردهای مهم دسته‌بندی تصویر، دسته‌بندی تصاویر با تعداد زیاد می‌باشد. به عنوان نمونه زمانی که بخواهیم در میان تعداد بسیار زیادی تصویر، یک تصویر معین یا همه‌ی تصاویر مربوط به یک موضوع خاص را بیابیم، بهترین راه استفاده از این تکنیک‌های دسته‌بندی است. یکی دیگر از کاربردهای دسته‌بندی تصاویر، تشخیص و تعیین نواحی در تصاویر ماهواره‌ای می‌باشد. تصاویر ماهواره‌ای مساحت بسیار زیادی را پوشش می‌دهند و از حجم بالایی از اطلاعات در تصویر برخوردارند. لذا برای بررسی این تصاویر و استخراج اطلاعات، آنها را به تعداد زیادی قطعات کوچک تقسیم می‌کنند و از تکنیک دسته‌بندی تصاویر کمک می‌گیرند.

۱ - ۲ هدف و فرضیات رساله

اگر بخواهیم مسئله دسته‌بندی تصویر را به‌طور دقیق‌تر بیان نمائیم، فرض کنید $X = \{x_1, x_2, \dots, x_N\}$ مجموعه‌ی تمام تصاویر آموزشی و تعداد این تصاویر $|X| = N$ باشد. مجموعه $Y = \{y_1, y_2, \dots, y_N\}$ نیز با همین تعداد اعضاء به عنوان برچسب‌های تصاویر در اختیار است. اگر کل داده‌ها به C دسته مختلف تعلق داشته باشند، آنگاه مقادیر $y_i \in \{1,2, \dots, C\}$ است. سیستم باید بتواند با بیشترین دقت برچسب داده‌های آزمون را مشخص کند. در این رساله فرض می‌شود که تصاویر استفاده شده شامل چندین شی می‌باشند و باید رابطه بین اشیاء موجود با استفاده از روش‌های تحلیل مدل‌های موضوعی استخراج شوند. در گام بعدی دسته‌بندی تصویر بر اساس محل قرار گرفتن اشیاء، برچسب آنها و مدل‌های موضوعی استخراج شده از تصویر انجام می‌گیرد. در مدل‌سازی موضوعی، فرض می‌شود که مجموعه تصاویر ورودی از روی چند موضوع نامعلوم ساخته شده است و هدف تشخیص این موضوعات می‌باشد. هر موضوع یک توزیع احتمال نامعلوم روی اشیاء موجود در تصویر است و هر تصویر توزیع احتمالی روی موضوع‌ها می‌باشد.

در یک دیدگاه می‌توان برای جستجوی یک تصویر از روش‌های مبتنی بر ظاهر یک شی استفاده کرد. در واقع می‌توان هنگام انجام عمل جستجو تمامی تصاویری که حاوی یک یا چند شی خاص می‌باشند را به عنوان نتیجه در نظر گرفت. اگرچه این تصاویر دربردارنده‌ی اشیاء مشابه هستند ولی ممکن است متعلق به موضوعات متفاوتی نیز باشند (شامل چند موضوع محلی) درحالی که هدف از دسته‌بندی فقط یک موضوع سراسری برای تصویر می‌باشد. برای انجام دقیق‌تر عمل جستجو، لازم است ابتدا موضوع مورد علاقه را در بین تصاویر جستجو و سپس جستجو را محدود به این تصاویر جدید کرد. در تصاویر جدید دوباره موضوع موردنظر محدودتر می‌شود و همین مراحل تکرار می‌شود تا جایی که تصاویر مورد نیاز شناسایی شوند.

به عنوان مثال، می‌خواهیم در آرشیو یک روزنامه موضوع خاص را جستجو نمائیم. موضوعات موجود در آرشیو شامل دسته‌های سیاسی، اقتصادی، فرهنگی، ورزشی و حوادث می‌باشند. هدف از جستجو

تشخیص موضوع سیاسی می‌باشد. موضوع سیاسی، شامل زیرموضوع‌های سیاست داخلی و خارجی می‌باشد که یکی از این زیرموضوعات انتخاب می‌شوند. همین مرحله دوباره تکرار می‌شود تا دقیقاً اسناد مورد نیاز شناسایی شوند.

لازم به ذکر است که این عمل برای تصاویر به‌سادگی امکان‌پذیر نیست؛ زیرا هر چه حجم تصاویر و اطلاعات افزایش می‌یابد دسته‌بندی فوق برای انسان کار مشکل و یا غیرممکنی می‌شود؛ در همین راستا باید از تکنیک‌های یادگیری ماشین استفاده شود. پژوهش‌گران حوزه‌ی یادگیری ماشین برای این عمل، مجموعه‌ای از الگوریتم‌ها تحت عنوان مدل‌سازی موضوعی آماری را توسعه داده‌اند.

الگوریتم‌های مدل‌سازی موضوعی روش‌های آماری هستند که اشیاء داخل یک تصویر را تحلیل کرده و از این طریق موضوعات داخل تصاویر را استخراج می‌کنند. این الگوریتم‌ها ارتباط این موضوعات با یکدیگر را مشخص می‌کنند. الگوریتم‌های مدل‌سازی موضوعی این امکان را فراهم می‌کنند تا سازمان‌دهی و خلاصه‌سازی آرشیوهای الکترونیکی را در ابعادی که از عهده‌ی انسان برنمی‌آید با دقت مناسبی انجام داد.

در مدل‌سازی موضوعی سه هدف زیر دنبال می‌شود:

- پیدا کردن موضوعات نامعلوم که در مجموعه تصاویر وجود دارند.
- تفسیر کردن تصاویر بر اساس موضوعات آن‌ها.
- بکارگیری این تفاسیر برای سازمان‌دهی، خلاصه‌سازی و جستجو تصاویر.

استخراج مدل‌های موضوعی پیش‌تر برای اسناد و کلمات موجود در هر سند معرفی شده است . بدست آوردن این موضوعات در تصویر با چالش اساسی مانند تشخیص اشیاء موجود در تصویر (همان کلمات یا نواحی که می‌توانند شامل کلمات باشند) و بدست آوردن نواحی مرتبط به این اشیاء می‌باشد. از مهمترین چالش‌های موجود در تشخیص اشیاء، وجود تصاویری از اشیاء با زاویه و نورپردازی مختلف و همچنین وجود هم‌پوشانی جزئی در اشیای داخل تصویر است. در واقع استخراج ویژگی در نواحی مرتبط به اشیاء می‌تواند در بالا بردن دقت سیستم‌های دسته‌بندی بسیار حائز اهمیت باشد. در این راستا، در

این رساله به منظور استخراج شی از تصاویر با وجود چالش‌های مطرح شده از روش‌های قدرتمند بر پایه شبکه‌های عصبی عمیق استفاده شده است. به منظور استخراج اشیاء، از روش‌های نواحی کاندیدا استفاده می‌شود در واقع بجای جستجو در کل تصویر از نواحی که می‌توانند به عنوان شی باشند استفاده می‌شود. یک مرحله مهم در گام بعدی، استخراج موضوعات نهفته در تصویر بر اساس اشیاء و نواحی مرتبط به آنها می‌باشد. در واقع در این فاز با توجه به این نواحی، کلمات بصری که می‌توانند مانند کلمات در متون باشند استخراج می‌شود. در ادامه از تکنیک‌های مدل‌سازی موضوعی برای استخراج موضوع‌های نهفته در تصویر استفاد می‌شود. در نهایت توسط این ویژگی‌ها و مدل‌های موضوعی عمل دسته‌بندی در تصویر انجام می‌شود.

۱ - ۳ چالش‌های رساله

دسته‌بندی تصویر به دلیل تنوع پس‌زمینه، زاویه دید متفاوت از اشیاء، تغییرات شدت روشنایی تصویر، هم‌پوشانی جزئی و کلی تصاویر یکی از دشوارترین مسائل پردازش تصویر می‌باشد. شکل ۱-۱ نمونه‌هایی از این پیچیدگی‌ها را نشان می‌دهد. همانطور که از این شکل مشخص است تصاویر استفاده شده در این رساله بسیاری از چالش‌های ممکن در تصویر را دارا است و هیچ محدودیتی از لحاظ پیچیدگی ظاهری برای تصاویر استفاده شده در نظر گرفته نشده است.

به طور کلی سیستم‌های دسته‌بندی تصاویر اغلب از دو بخش اساسی استخراج ویژگی و طبقه‌بند تشکیل می‌شوند. در بخش استخراج ویژگی به ازای هر تصویر بردار ویژگی استخراج می‌شود. سپس این بردارهای ویژگی برای آموزش و آزمون طبقه‌بند استفاده می‌شوند. پیچیدگی‌های موجود در تصویر موجب ایجاد تغییرات غیرخطی در فضای ویژگی نمونه‌های یک دسته می‌شود که جداسازی در فضای ویژگی را

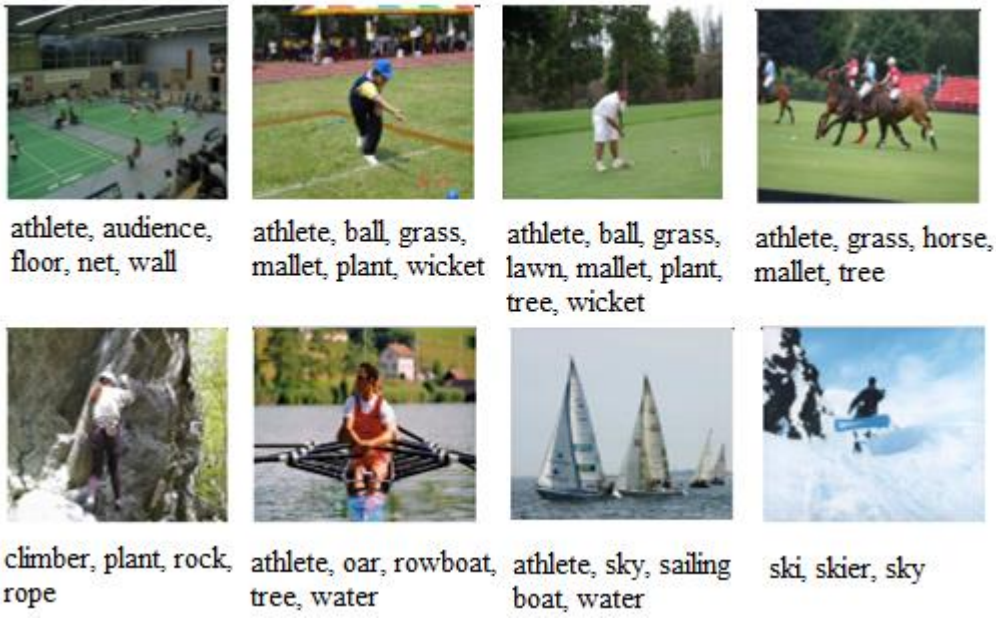
شکل ۱-۱ نمونه‌هایی از پیچیدگی‌های موجود در دسته‌بندی صحنه

ناکارآمد می‌سازد. بنابراین با در نظر گرفتن پیچیدگی‌های تصویر در بخش استخراج ویژگی، عمل دسته‌بندی با دقت بیشتری انجام خواهد شد. در واقع محتوای بصری تصاویر به صورت ویژگی‌های سطح پایین مانند رنگ و بافت می‌باشند و ویژگی‌های معنایی به صورت ویژگی‌های سطح بالایی مانند اشیاء و برچسب آنها می‌باشد. این دو دسته‌ی ویژگی سطح پایین و سطح بالا فاصله زیادی باهم دارند که کشف رابطه‌ی بین آنها کار دشواری است. به این فاصله، فاصله معنایی[۱] گفته می‌شود. باید این فاصله معنایی با استخراج ویژگی‌های بهتر و برچسب‌زنی به تصاویر کاهش یابد. در واقع بدست آوردن ویژگی‌های معنایی بهتر می‌تواند در بالا بردن دقت الگوریتم‌های دسته‌بندی حائز اهمیت باشد.

[۱] Semantic Gap

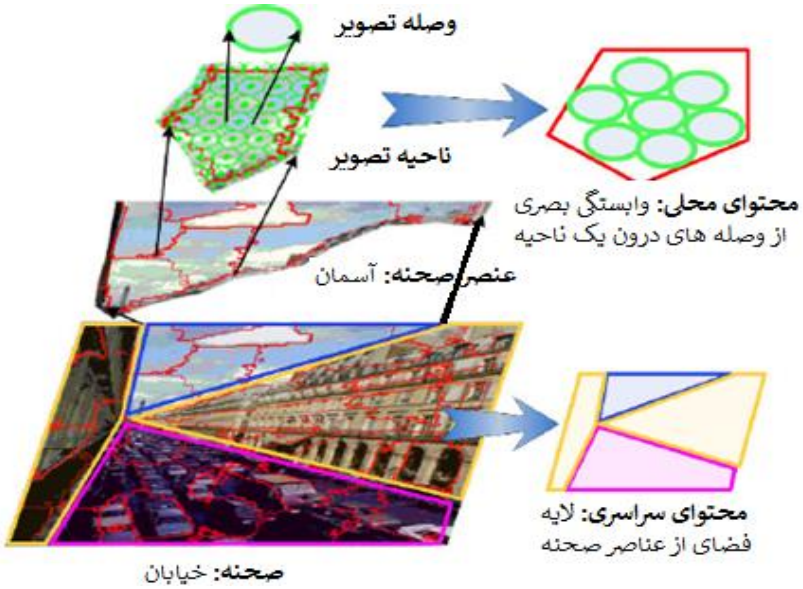
دسته‌بندی و برچسب‌زنی تصویر دو وظیفه مهم در بینایی ماشین است. هدف از دسته‌بندی تصویر، توصیف سراسری با یک برچسب توصیف‌کننده است. ساحل، فضای باز، درون شهر و غیره نمونه‌هایی از برچسب‌های یک تصویر هستند. برچسب‌زنی تصویر، به برچسب‌گذاری محتوایی محلی یک تصویر تاکید دارد. برای نمونه، یک تصویر می‌تواند شامل، "آسمان"، "ماشین"، "درخت" و غیره باشد. این دو مسئله به هم مرتبط هستند. پس تلاش برای حل مشترک این مسائل طبیعی است. برای نمونه در یک تصویر با برچسب خیابان، احتمال توصیف تصویر با حضور "ماشین"، "انسان" یا "ساختمان" بیشتر از "ساحل" یا "آب دریا" است. کارهای زیادی بر روی دسته‌بندی و برچسب‌زنی تصویر به صورت جداگانه انجام شده است. ولی کارهای بسیار کمی برای حل همزمان این دو مشکل انجام شده است.

یک چالش مهم در این نوع از سیستم‌ها تشخیص اشیاء یا محتوای درون تصویر می‌باشد. در سال‌های اخیر روش‌های بسیاری برای تشخیص و برچسب‌زنی اشیاء معرفی شده است ولی همچنان نتایج برای تشخیص اشیاء با تعداد زیاد قابل قبول نمی‌باشد. در این نوع از سیستم‌ها یک مرحله مهم بدست آوردن نواحی کاندیدایی می‌باشند که می‌توانند به عنوان شی معرفی شوند. در واقع در فاز بعدی برای تشخیص اشیاء، بجای جستجو در کل تصویر فقط تمرکز بر روی نواحی کاندیدا خواهد بود. تشخیص و برچسب‌زنی بسیار وابسته به این نواحی می‌باشد. شکل ۱-۲ نمونه‌هایی از تصاویر موجود در یک مجموعه از تصاویر ورزشی با اشیاء موجود در آنها را نشان می‌دهد. همانطور که در شکل ۱-۲ نشان داده شده است، تصاویر مرتبط با ورزش‌های دریایی شامل محتواهایی مانند "آب" و "قایق" و تصاویری که در فضای‌های آزاد انجام می‌شوند شامل محتوایی مانند "چمن" می‌باشند. در واقع در این نوع از تصاویر، استخراج اشیاء موجود و بدست آوردن موضوعات پنهان در آنها می‌تواند در بالابردن دقت عملکرد دسته‌بند تاثیرگذار باشد. بدست آوردن این موضوعات وابسته به نواحی/محل قرار گرفتن اشیاء و برچسب اشیاء می‌باشد. در ادامه می‌توان بر اساس این خصیصه‌ها مفاهیم پایه‌ای مانند کلمات بصری را مناسب‌تر و دقیق‌تر استخراج کرد. ولی معمولا به دلیل تعدد اشیاء، پس‌زمینه و هم‌پوشانی جزئی بین اشیاء و پس‌زمینه این امر ساده نمی‌باشد. در این رساله راهکاری برای غلبه بر این چالش‌ها مطرح می‌شود.



شکل ۱-۲ نمونه‌هایی از تصاویر ورزشی با محتوا و برچسب‌های متفاوت [۱]

شکل ۱-۳ نمونه ای از مدل‌سازی با چهار سطح از اطلاعات بصری در صحنه‌ی تصویر را نشان می‌دهد. در این تصویر، دسته صحنه "خیابان"، عناصر صحنه "آسمان"،"ساختمان" و "جاده" می‌باشد. نواحی موجود در تصویر با محدوده قرمز نشان داده شده است. تکه‌های تصویر با دایره سبز رنگ نشان داده شده است. محتوای سراسری نشان دهنده لایه فضایی از عناصر صحنه می‌باشد. محتوای محلی نشان دهنده رابطه بین تکه‌های‌های تصویر درون ناحیه نصویر می‌باشد.



شکل ۱-۳ مدل‌سازی با چهار سطح از اطلاعات بصری در صحنه‌ی تصویر

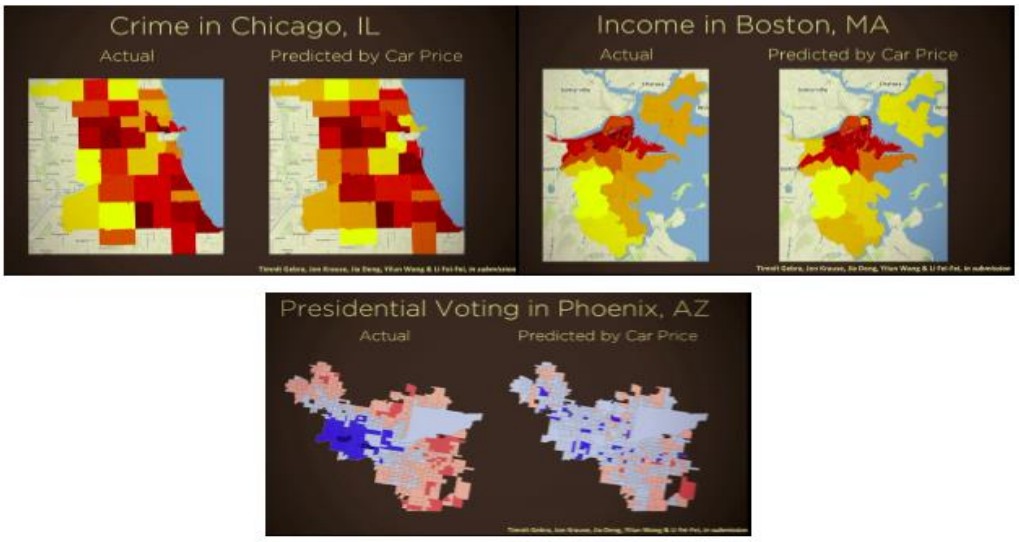
در واقع در مدل‌سازی موضوعی، اشیاء درون یک تصویر را تحلیل می‌کنیم و از طریق آنها موضوعات نهفته در تصاویر را استخراج می‌کنیم. با استفاده از این روش ارتباط موضوعات با یکدیگر مشخص می‌شود. هدف از این مرحله، پیدا کردن موضوعات نامعلوم موجود در مجموعه تصاویر و تحلیل تصاویر بر اساس موضوعات آن‌ها می‌باشد که این عمل، دسته‌بندی و جستجو را با دقت بیشتر امکان‌پذیر می‌سازد.

۱ - ۴ کاربرد دسته‌بندی موضوعی تصویر

توصیف خودکار محتوای تصاویر به عنوان یکی از مشکلات اساسی در هوش مصنوعی مطرح بوده است که این امر سبب ارتباط بیشتر بین بینایی ماشین و پردازش زبان طبیعی شده است. به همین دلیل، در سال‌های اخیر محققان بسیاری به مطالعه درباره این موضوع پرداخته‌اند. در حال حاضر توصیف خودکار تصویر و دسته‌بندی آن بر اساس محتوا یکی از چالش‌های بحث برانگیز در بین محققان است [۱, ۲]. بهبود این نوع از سیستم‌ها، می‌تواند تأثیر بسیاری در سیستم‌های تحت وب و موتورهای جستجو داشته باشد. برای نمونه، این نوع سیستم‌ها می‌توانند در پی‌بردن به محتوای تصاویر موجود در وب و بهبود دقت جستجو، برای مردمی با اختلالات بینایی، کمک شایانی کنند. در واقع در این نوع از سیستم‌ها علاوه بر استخراج اشیاء موجود در تصاویر باید چگونگی ارتباط بین این اشیاء نیز استخراج شوند. همانطور که اشاره شد، یکی از کاربردهای مهم این نوع از سیستم‌ها در موتورهای جستجو است. در بسیاری از کاربردها این نیاز احساس می‌شود که یک تصویر بر اساس محتوا، در دسته‌ای مشخص قرار گیرد و با توجه به آن، عمل جستجو با دقت و سرعت بیشتری انجام شود. معمولاً این نوع جستجوها، در حجم بالایی از داده‌های وب هستند. معمولاً در این نوع از سیستم‌ها تنوع زیادی در حالت‌ها و موقعیت‌های اشیاء وجود دارد. در همین راستا باید از تکنیک‌های استفاده شود که با حجم بالایی از داده در ارتباط باشند و بتوانند مدل‌های مختلفی از یک شی را آموزش دهند.

یکی دیگر از کاربردهای مهم این نوع از سیستم‌ها، فیلتر نمودن تصاویر غیراخلاقی یا سانسور افراد یا محیط‌های خاص در نتایج جستجو است. در خیلی از موارد، خانواده‌ها نمی‌خواهند افراد کم سن به مجموعه‌ای از تصاویر با توجه به شرایط فرهنگی دسترسی داشته باشند و با این روش‌ها می‌توان به تحلیل این نوع از تصاویر پرداخت و تصاویر نامناسب را از دسترس افراد غیرمجاز خارج نمود. البته باید به این نکته دقت نمود که میزان نامناسب بودن تصاویر در زمینه‌های مختلف متفاوت بوده است. برای نمونه تعریف تصاویر نامناسب در سامانه‌ای مانند آپارات با رسانه ملی متفاوت است. در واقع میزان این نامناسب بودن در کاربردهای مختلف با توجه به نیاز کاربران با هم فرق می‌کند که باید به تحلیل این نوع از مسائل نیز پرداخت. در گام بعدی باید با توجه به محتوای تصویر و اشیاء موجود در آن، این نیازها را برطرف نمود.

یکی دیگر از کاربردهای مهم بررسی محتوای تصویر بر اساس اشیاء، تحلیل رفتاری افراد است. برای نمونه می‌توان به کاری در دانشگاه استنفورد اشاره نمود. نتایج این تحقیق در شکل ۱-۴ نشان داده شده است. در این تحقیق با بررسی میلیون‌ها تصویر "منظره خیابان بدست آمده از گوگل" در صدها شهر آمریکا نتایج جالبی بدست آمد: اول اینکه قیمت خودرو وابستگی زیادی به درآمد خانوارها دارد، همچنین قیمت خودرو بستگی زیادی هم به نرخ جرایم در شهرها دارد. همچنین در تحقیق دیگری در این دانشگاه، نشان داده شده که الگوی رأی دادن افراد در شهرها به موقعیت جغرافیایی افراد (کد پستی) وابسته بوده است.



شکل ۱-۴ نتایج تحلیلی از آزمایشات بر روی تصاویر و برداشت‌های بدست آمده با توجه به موقعیت جغرافیایی [۳]

یکی دیگر از تحلیل‌های این نوع از داده‌ها، بررسی تغییر نوع تصاویر استفاده شده در شبکه‌های اجتماعی است. با این تحلیل، می‌توان رفتار افراد را با توجه به تصاویر استفاده شده در پروفایل شخصی آن‌ها در این نوع از شبکه‌ها ارزیابی نمود. برای مثال می‌توان بررسی کرد که افراد در قسمت‌های مختلف جغرافیایی چه نوع تصاویر و چه اشیایی را بغیر از تصویر خود، در پروفایل‌های شخصی استفاده می‌کنند. (استفاده از تصویر سگ در بسیاری موارد برای کاربران ایرانی در شبکه‌های اجتماعی مانند Facebook نشانه غرب‌زدگی است.) برای بدست آوردن این اطلاعات به سیستم قوی تحلیل محتوای تصاویر نیاز است. با توجه به این اطلاعات، می‌توان رفتارهای افراد در جوامع مختلف را با توجه به شرایط فرهنگی مورد تحلیل قرار داد.

۱ - ۵ دستاوردهای رساله

در تحقیقات انجام شده در سال‌های اخیر، بدست آوردن ناحیه‌های کاندیدا به عنوان یک مرحله اساسی و مهم در سیستم‌های تشخیص و شناسایی اشیای موجود در تصویر مطرح می‌باشد. بدست آوردن این ناحیه‌ها مانند یک تنگنا بوده و بیشترین بار محاسباتی را در این نوع از سیستم‌ها دارد. در همین راستا انتخاب روش مناسب و سریع به منظور پیشنهاد نواحی کاندیدا، می‌تواند در بهبود عملکرد سیستم‌های تشخیص بسیار حائز اهمیت باشد. در مقاله زیر، به مرور کارهای انجام شده در این زمینه پرداخته شده

است و چندین روش مشهور و محبوب در سیستم‌های شناسایی قدرتمند معرفی شده است. همچنین در این مقاله به مقایسه و ارزیابی روش‌های مطرح بر روی مجموعه‌داده‌های استاندارد موجود پرداخته شده است. با توجه به نتایج بدست آمده در این مقاله شبکه‌های عمیق بهترین عملکرد را در این نوع از سیستم‌ها از خود نشان داده است.

علی. قنبری سرخی ، حمید. حسن‌پور و منصور. فاتح، "*انتخاب ناحیه‌های کاندیدا در سیستم‌های تشخیص و شناسایی اشیاء*"، مجله محاسبات نرم، جلد ۵، شماره ۲، صفحات ۳۴-۴۷، ۱۳۹۵.

با توجه به نتایج بدست آمده از مقاله قبل، استفاده از نواحی کاندیدا می‌تواند در بالا بردن عملکرد سیستم‌های دسته‌بندی تصویر بسیار حائز اهمیت باشد. در سال‌های نه چندان دور از روش‌های پایه یادگیری ماشین برای دسته‌بندی تصویر استفاده شده است. در بسیاری از کارهای انجام شده در سال‌های بین ۲۰۰۰ تا ۲۰۰۶ از ترکیبی از چند روش برای دسته‌بندی تصویر استفاده شده است. ولی با توجه به نتایج گزارش شده، این روش‌ها فقط برای مجموعه داده‌های کوچک با پیچیدگی کم می‌تواند عملکرد خوبی داشته باشد. در مقاله زیر، یک راهکار نوین از ترکیب روش‌های سنتی یادگیری ماشین به منظور دسته‌بندی تصویر با تعداد دسته کم ارائه شده است. در این مقاله از نواحی کاندیدا برای استخراج نواحی، از روش پیچش و تجزیه مقادیر منفرد برای استخراج ویژگی، از شبکه Boosting برای شناسایی برچسب نواحی و از ماشین بردار پشتیبان برای دسته‌بندی تصویر استفاده شده است. نتایج بدست آمده از روش پیشنهادی بر روی مجموعه داده SUN[1] نشان دهنده عملکرد خوب این روش بر روی مجموعه‌داده‌هایی با تعداد کلاس‌های کم می‌باشد. ولی چالش اساسی در این روش‌ها بار محاسباتی بالای این روش‌ها (زمان اجرای مرحله آموزش و آزمون) می‌باشد. در واقع از نتایج بدست آمده از این مقاله می‌توان به این نکته مهم رسید که این روش‌ها نمی‌توانند در سیستم‌های واقعی دسته‌بندی تصویر نتایج خوبی را از خود نشان دهند.

[1] Scene UNderstanding (SUN) database

علی. قنبری سرخی ، حمید. حسن‌پور و منصور. فاتح، "*دسته‌بندی صحنه تصویر بر پایه روش ناحیه کاندیدا و شبکه‌های Boosting*"، اولین کنگره و نمایشگاه بین‌المللی علوم و تکنولوژی‌های نوین، بابل، ایران، ۱۳۹۷.

با توجه به مطالعات انجام شده در این رساله، می‌توان دسته‌بندی تصویر را به دو دسته کلی تصاویر با پیچیدگی کم و تعداد کلاس‌های پایین و تصاویر با پیچیدگی بالا و تعداد کلاس‌های بالا بخش بندی نمود. در همین راستا با توجه به مجموعه داده‌های موجود، در قسمتی از این رساله، به معرفی روشی نوین برای دسته‌بندی تصاویر نامتعارف پرداخته شده است. این نوع از تصاویر دو کلاسه شامل تصاویر نامتعارف و متعارف می‌باشند.

یکی از کاربردهای مهم دسته‌بندی تصویر فیلتر کردن تصاویر و تشخیص تصاویر نامتعارف می‌باشد. در واقع در سال‌های اخیر با پیشرفت روزافزون اینترنت و رسانه‌های تحت وب، توزیع و اشتراک منابع اطلاعاتی نظیر تصویر در حال افزایش است. اشتراک این منابع علاوه بر مزایای بسیار، خطرات و مشکلاتی نظیر دسترسی به تصاویر نامتعارف دارد که به نوبه‌ی خود تهدیدی برای فرهنگ جوامع مختلف، به‌خصوص نوجوانان و جوانان است. به همین منظور، در مقاله زیر، به تحلیل و بررسی روشی برای دسته‌بندی تصاویر نامتعارف و فیلترینگ هوشمند آن‌ها پرداخته شده است. یکی از مشکلات این نوع از سیستم‌ها، حجم بالای داده‌های موجود در شبکه‌های تحت وب و استخراج ویژگی‌های معنادار در این حجم از داده‌ها است. در این راستا، در این مقاله روشی جدید، بر پایه شبکه‌های عصبی عمیق به منظور تشخیص هوشمند تصاویر نامتعارف ارائه شده است. در واقع یک معماری نوین برای دسته‌بندی پیشنهاد شده است. نتایج بدست آمده نشان‌دهنده بهبود عملکرد روش پیشنهادی بر روی مجموعه داده‌های به‌نسبت بزرگ می‌باشد.

علی. قنبری سرخی ، منصور. فاتح و حمید. حسن‌پور، "*تشخیص و فیلترینگ هوشمند تصاویر نامتعارف به کمک شبکه‌های عصبی عمیق*"، فصل‌نامه علمی-پژوهشی پردازش علائم و داده‌ها، جلد ۱۵، شماره ۲، صفحات ۵۵-۶۸، ۱۳۹۷.

در ادامه تمرکز این رساله ارائه روشی برای بهبود دسته‌بندی تصاویر با پیچیدگی بیشتر می‌باشد. همانطور که اشاره شد، امروزه آشکارسازی و برچسب‌زنی اشیاء در تصاویر یکی از چالش‌های اساسی در

سیستم‌های دسته‌بندی تصویر می‌باشد. در سال‌های اخیر استفاده از یادگیری عمیق مورد توجه محققان قرار گرفته است. در همین راستا، در مقاله زیر، ابتدا جدیدترین شبکه‌های عمیق معرفی شده و نقاط قوت و ضعف آنها تحلیل شده است. در ادامه شبکه‌ای بهبود یافته از شبکه R-FCN ارائه شده است. روش پیشنهادی بر پایه معماری ResNet و شبکه تمام پیچش است. در این روش، معماری جدیدی مبتی بر شبکه عمیق برای پیشنهاد ناحیه کاندیدا و روشی ترکیبی مبتنی بر SVMفازی دوکلاسه و SVR برای آشکارسازی و برچسب‌زنی اشیاء مرتبط به مجموعه بصری ارائه شده است. در این روش از تابع زیان جدید با عنوان اختلاف کوشی-شوارتز استفاده شده است. این تابع زیان از لحاظ سرعت و دقت، عملکرد بهتری از خود نشان داده است. نتایج بدست آمده نشان دهنده بهبود عملکرد (از لحاظ دقت و زمان اجرا مرحله آزمون) این روش نسبت به روش پایه شبکه R-FCN است. با توجه به نتایج بدست آمده می‌توان بیان نمود که معماری تمام پیچش نسبت به معماری‌های دیگر شبکه‌های عمیق عملکرد بهتری در آشکارسازی اشیاء از خود نشان می‌دهند.

علی. قنبری سرخی ، حمید. حسن‌پور و منصور. فاتح، "*بهبود شبکه عمیق R-FCN در آشکارسازی و برچسب‌زنی اشیاء*"، مجله ماشین بینایی و پردازش تصویر (پذیرش).

در سال‌های اخیر بدست آوردن موضوعات مهم در متن و تصویر برای دسته‌بندی بر اساس موضوعات پنهان مورد توجه محققان بسیاری قرار گرفته شده است. در همین راستا در مقاله زیر، (در ادامه مقاله فوق)، با توجه به ویژگی‌های سطح بالای استخراج شده از شبکه تمام پیچش، نواحی و برچسب استخراج شده از مرحله قبل، موضوعات نهفته در تصویر استخراج شده است. در سیستم‌های که تاکنون برای بدست آوردن موضوعات پنهان معرفی شده، برچسب‌ها از قبل از آماده بوده است ولی در روش پیشنهادی از برچسب‌ها و اشیاء استخراج شده از مرحله قبل به عنوان ورودی برای بدست آوردن موضوعات پنهان استفاده شده است. در روش ارائه شده، از یک روش نوین برای استخراج مجموعه کلمات ارائه شده است. در مرحله بعد روش‌های استخراج موضوعات پنهان مورد ارزیابی قرار گرفته شده است. در این مقاله از SupDocNADE با یک رویکرد جدید استفاده شده است. در این روش از لایه‌ی میانی مرتبط به

موضوعات پنهان استفاده شده است. در نهایت با ترکیب ویژگی‌های استخراج شده از برچسب برای کلمات بصری، پیچش و موضوعات پنهان دسته‌بندی توسط یک بیزین انجام شده است. نتایج نشان دهنده بهبود عملکرد (از لحاظ دقت و زمان اجرای مرحله آزمون) روش پیشنهادی در مقایسه با روش‌های دیگر مبتنی بر شبکه‌های عمیق می‌باشد.

A. Ghanbari sorkhi, H. Hassanpour and M. Fateh, "*A Comprehensive System for Image Scene Classication*", Multimedia Tools and Applications. (Submitted in November 2018)

۱ - ۶ سازماندهی طرح تحقیق

همانطور که در بخش‌های مختلف این فصل بیان شد، طراحی یک سیستم یکپارچه به منظور دسته‌بندی صحنه با پیچیدگی بالا شامل چندین بخش می‌باشد. بدست آوردن کلمات بصری و استخراج ویژگی‌های سطح بالا نقش مهمی در استخراج مدل‌های موضوعی در تصویر دارد. در واقع بدست آوردن این کلمات با محدودیت‌های مانند مکان نواحی مرتبط به ویژگی بصری، بدست آوردن نماینده هر کلمه بصری، میزان شباهت هر ویژگی بصری با نماینده، پیچیدگی زمانی و مکانی روبرو می‌باشد. یکی دیگر از بخش‌های مهم در این نوع از سیستم‌ها بدست آوردن نواحی مرتبط به اشیاء و برچسب آنها می‌باشد که مرحله بدست آوردن مدل موضوعی به این دو بخش بسیار وابسته می‌باشد در همین راستا در این رساله بخش‌های مختلف از سیستم دسته‌بندی تصویر تشریح می‌شود و به تفصیل روش پیشنهادی برای طراحی یک سیستم یکپارچه توضیح داده می‌شود. این رساله در ۶ فصل تهیه شده است. هر فصل به شرح زیر است.

۱. فصل دوم بر کارهای انجام شده در زمینه دسته‌بندی تصویر تمرکز دارد. در ابتدا روش‌های استخراج ویژگی و روش‌های استخراج کلمات بصری معرفی شده است.در گام بعدی استخراج

نواحی کاندیدا معرفی شده در آخر روش‌های موجود در دسته‌بندی تصاویر نامتعارف، تصاویر با صحنه‌های پیچیده معرفی شده است.

۲. در فصل سوم روش پیشنهادی در دسته‌بندی صحنه معرفی شده است. در ابتدا به معرفی روش‌پیشنهادی در تصاویر با پیچیدگی کم مانند تصاویر نامتعارف معرفی شده است. در ادامه فصل تمرکز بر دسته‌بندی تصاویر با پیچیدگی بالا بر اساس موضوعات نهفته می‌باشد. روش پیشنهادی در این فصل معرفی شده است.

۳. در فصل چهارم پیاده‌سازی و آزمایشات انجام شده بر اساس روش‌های پیشنهادی در این رساله ارائه شده است.

۴. در فصل پنجم نتیجه‌گیری و کارهای آتی در موضوع دسته‌بندی تصویر بیان شده است.

۱ - ۷ جمع‌بندی

در سال‌های اخیر با افزایش روز افزون تصاویر دیجیتال، طراحی یک سیستم خودکار دسته‌بندی تصویر بر اساس محتوا به عنوان یک کاربرد مهم در بینایی ماشین و پردازش تصویر است. در همین راستا در این فصل، به بیان و تعریف مسئله دسته بندی صحنه تصویر پرداخته شده است. سپس چالش‌ها، فرضیات و هدف اصلی رساله بیان شد. دستاوردهای رساله با توجه به چالش‌های موجود به صورت کلی بیان شد. همچنین در انتهای این فصل سازماندهی کلی رساله به اختصار تشریح شد. .

۲ پیشینه تحقیق

۲ - ۱ مقدمه

در این فصل به معرفی کارهای انجام شده برای دسته‌بندی تصاویر پرداخته شده است. روش‌های موجود در این کاربرد شامل بخش‌های مختلفی می‌باشد. این بخش‌ها شامل استخراج ویژگی معنایی مانند مجموعه کلمات[۱] استخراج ویژگی‌های سطح بالا مانند نواحی مرتبط به شی، برچسب اشیاء و روش‌های موجود برای دسته‌بندی بر اساس محتوا می‌باشد. در همین راستا، تحقیقات انجام شده در این کاربرد بر اساس بخش‌های مختلف در ادامه معرفی می‌شوند.

درک صحنه بصری به عنوان یکی از موضوعات فعال در تحقیقات انجام شده در بینایی ماشین است [۴]. مشخصه‌های ظاهری محلی تصویر اغلب نشانه‌های[۲] بصری با ارزشی را برای شناسایی صحنه فراهم می‌کند. اکثر این روش‌ها، پرورش یافته‌ی مدل مجموعه کلمات هستند. در گام نخست، ویژگی‌های محلی مستخرج از تصویر، به یک مجموعه از کلمات بصری[۳] نگاشت می‌شود. این نگاشت توسط بردار رقمی‌سازی[۴] انجام می‌شود. در گام بعدی، یک تصویر از هیستوگرام رخدادهای کلمه بصری نمایش داده می‌شود که به صورت طبیعی، آماری از ویژگی‌های محلی کدگذاری[۵] شده است. روش‌های زیادی برای این کاربرد در سال‌های اخیر معرفی شده است که در ادامه به معرفی این روش‌ها پرداخته می‌شود.

۲ - ۲ استخراج مجموعه کلمات

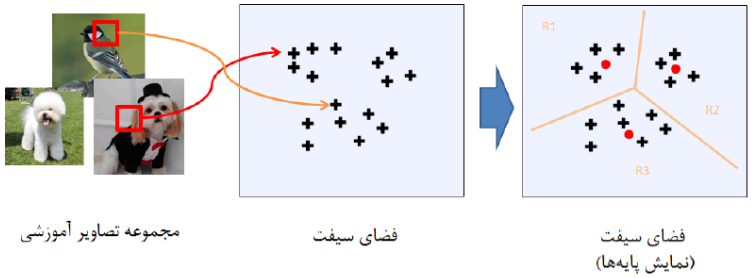
استفاده از نقاط محلی برای تولید ویژگی، با الهام از روش مجموعه کلمات اولین بار در سال ۲۰۰۴ در مرجع [۵] مطرح شد و تا امروز نسخه‌های متنوعی از آن ارائه شده است. در مرحله اول، این روش تعدادی نقطه به همراه همسایگی مشخصی از هر تصویر انتخاب می‌شود که به آن‌ها نقاط محلی

---

[۱] Bag-of-Words (BoW)
[۲] Cue
[۳] Visual Words
[۴] Quantization
[۵] Encodes

می‌گویند. در مرحله بعد، تمامی این نواحی محلی با روش یکسان‌سازی هیستوگرام، از نظر شدت نور و اندازه، یکسان‌سازی می‌شوند. بدین ترتیب نواحی محلی نسبت به تغییرات نور و اندازه، تا حد خوبی مستقل می‌شوند. سپس از این نقاط، ویژگی SIFT[1] استخراج می‌شود. ویژگی SIFT، هیستوگرامی از اندازه و جهت بردارهای گرادیان تصویر در نقاط تصویر است. این ویژگی نسبت به دوران تصویر مستقل است. خروجی این الگوریتم یک بردار ۱۲۸ بعدی ویژگی محلی است که نسبت به تغییرات شدت نور، اندازه و دوران تا حد خوبی مستقل است. در اینجا، یک بردار ویژگی SIFT به‌صورت S_i نشان داده می‌شود که به معنی ویژگی SIFT استخراج‌شده از i–امین نقطه محلی تصاویر است.

در مرحله دوم، خوشه‌بندی بردارهای SIFT حاصل از تمامی تصاویر انجام می‌شود و مرکز هر خوشه به عنوان ویژگی سرآمد انتخاب می‌گردد. در واقع مرکز هر خوشه یک ویژگی سرآمد است که به بهترین شکل می‌تواند داده‌های موجود در یک خوشه را توصیف نماید. منظور از توصیف در روش‌های مختلف متفاوت است و می‌تواند به معنی کم‌ترین خطای بازسازی یا نزدیکی اقلیدسی به داده‌ها باشد. l–امین مرکز خوشه به‌صورت d_l نشان داده می‌شود. به هر مرکز خوشه به اصطلاح پایه و به مجموعه پایه‌ها واژه‌نامه گفته می‌شود. در شکل ۲-۱ مراحل استخراج ویژگی و خوشه‌بندی نمایش داده شده است.



شکل ۲-۱ مراحل یادگیری پایه‌ها در مجموعه کلمات بر اساس الگوریتم ارائه شده در مرجع [۵]

مرحله سوم به کدگذاری معروف است. در این مرحله به هر ویژگی SIFT، S_i یک کد a_i اختصاص داده می‌شود. ابعاد بردار کد a_i برابر با تعداد مراکز خوشه است و همه عناصر آن به جز یک عنصر برابر با صفر است. اگر فرض کنیم داده S_i به l-امین خوشه تعلق داشته باشد عنصر l-ام کد a_i برابر با یک و باقی

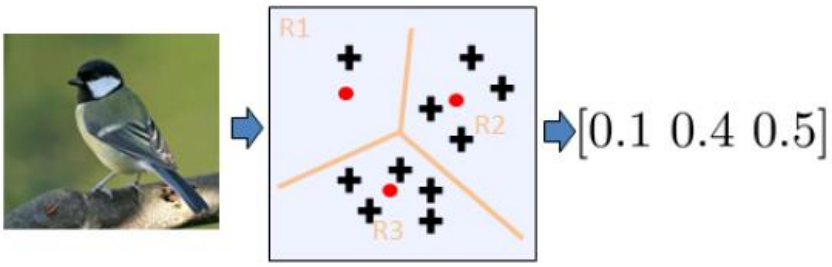
[1] Scale-Invariant Feature Transform (SIFT)

عناصر برابر با صفر قرار داده می‌شوند. با کمی تأمل مشخص می‌شود که درواقع بردار کد، یک نمایش سطح بالا از ویژگی‌های سطح پایین SIFT است. در بردار کد به جای استفاده از ویژگی سطح پایین، از ارتباط آن با ویژگی‌های شاخص دیگر برای نمایش آن استفاده شده است.

در مرحله چهارم، کدهای حاصل از هر تصویر به یک ویژگی کلیِ معرف تصویر اولیه تبدیل می‌شوند. در این مرحله تجمیع با عمل هیستوگرام انجام می‌گردد. به این ترتیب بردار ویژگی نهایی با رابطه (۲-۱) به دست می‌آید:

$$v_i = \frac{1}{m} \sum_{j \in image(i)} a_j \qquad (۲-۱)$$

که v_i بردار ویژگی نهایی تصویر h-ام و m تعداد نواحی محلی استخراج شده از این تصویر است. عمل هیستوگرام روی المان‌های بردار a_j انجام می‌شود و بنابراین ابعاد V_i با ابعاد بردار کد برابر است. در شکل ۲-۲ شیوه کدگذاری یک تصویر به شکل نمادین نشان داده شده است.



شکل ۲-۲ ایجاد بردار ویژگی در روش مجموعه کلمات بر اساس الگوریتم ارائه شده در مرجع [۵]

اغلب روش‌های استخراج ویژگی محلی شامل این مراحل هستند:

۱. انتخاب نواحی محلی تصویر
۲. استخراج ویژگی از نواحی محلی
۳. ساخت پایه‌ها یا ویژگی‌های سرآمد (یادگیری واژه‌نامه[۱])
۴. نمایش هر ویژگی بر اساس پایه‌ها (کدگذاری)
۵. ترکیب کد نواحی محلی برای محاسبه ویژگی نهایی (انتخاب[۲])

[۱] Dictionary Learning
[۲] Pooling

نحوه انتخاب نواحی محلی یکی از مسائل باز است. با عدم انتخاب نواحی با مفاهیم اصلی تصویر، کار طبقه‌بندی بسیار مشکل می‌شود. این نواحی معمولاً به طور تصادفی، بر روی لبه‌ها یا به طور یکنواخت از تصویر، انتخاب و استخراج می‌گردنند. آزمایش‌های بسیاری برای مقایسه این روش‌ها با یکدیگر انجام شده است [۶] که نتایج نشان می‌دهد که هیچ‌کدام از این روش‌ها به‌طور کلی بهتر از دیگری نیستند.

روش مطلوب در استخراج ویژگی، بردار ویژگی مستقل از چرخش تولید می‌کند. از طرفی لبه‌های تصویر، حاوی اطلاعات بیشتری نسبت به نواحی هموار تصویر هستند. از این‌رو استفاده از روش‌های مبتنی بر هیستوگرام بردار گرادیان[۱] مانند SIFT و هیستوگرام گرادیان، سودمند هستند و در استخراج ویژگی محلی از آن‌ها استفاده می‌شوند.

مشکل این الگوریتم در مرحله کدگذاری، خطای بازسازی[۲] بالای آن است. به طور دقیق‌تر، بازسازی ویژگی اولیه از ویژگی‌های توصیفی خطای زیادی دارد. برای حل این مشکل، ضمن ارائه چارچوب تئوری، پیشنهاد شده که هر داده با چندین پایه نزدیک به آن توصیف شود که این روش‌ها در بخش بعدی مورد بررسی قرار گرفته‌اند. مشکل دیگر این روش، عدم تأثیرگذاری مکان ویژگی‌ها درنتیجه نهایی است. یعنی با تغییر مکان نقاط محلی نسبت به هم، بردار ویژگی نهایی تغییر نمی‌کند. این مشکل را می‌توان در مرحله انتخاب، با لحاظ کردن مکان رویداد ویژگی‌ها در تصویر برطرف کرد. چالش دیگرِ مرحله انتخاب، عملگر استفاده‌شده در تجمیع کدها است. روش معمول در این بخش، استفاده از هیستوگرام است. اما به تازگی به‌طور شهودی مشاهده شده که استفاده از عملگر بیشینه‌سازی نتایج بهتری به همراه دارد [۷].

[۱] Histogram of Gradients (HoG)

[۲] Reconstruction Error

۲ - ۲ - ۱روش هرم تطابق مکانی

همان‌گونه که اشاره شد، یکی از ایرادهای روش مجموعه کلمات، استقلال ویژگی نهایی از مکان قرارگیری نقاط محلی در تصویر است. روش هرم تطابق مکانی[۱][۸] برای حل این مشکل پیشنهاد شده است. ایده اصلی این روش از روش هسته تطبیق هرمی [۹] الهام گرفته شده است. بنابراین قبل از معرفی روش هرم تطابق مکانی، روش هسته تطبیق هرمی بررسی می‌گردد.

۱) هسته تطبیق هرمی

این روش برای تشخیص شباهت میان دو مجموعه به کار می‌رود. فرض کنید X و Y مجموعه‌ای از بردارهای با بعد d باشند. هسته تطبیق هرمی، ابتدا یک شبکه منظم را در فضای برداری در نظر می‌گیرد. در ادامه، با شمارش تعداد داده‌های هر سلول شبکه از مجموعه X و Y، مقدار هسته در آن شبکه مشخص می‌شود. به طور دقیق‌تر، اگر شبکه‌هایی در سطوح 0،...،L داشته باشیم، هر بعد شبکه lام، به 2^l سلول یک اندازه تقسیم می‌شود و در نهایت $D=2^{dl}$ سلول داریم. فرض کنید $H_X^l(i)$ و $H_Y^l(i)$ به ترتیب تعداد داده‌های شبکه lام در سلول iام از مجموعه X و Y هستند. حال تعداد تطبیق‌های دو مجموعه در سطح lام با رابطه (۲-۲) نمایش داده می‌شود:

$$I_l = \sum_{i}^{D} f(H_Y^l(i), H_X^l(i)) \qquad (۲-۲)$$

در رابطه (۲-۲)، f تابع شمارش می‌باشد که تعداد تطبیق را محاسبه می‌کند و I_l تعداد تطبیق‌ها در سطح lام است. تعداد تطبیق‌ها در این سطح، تطبیق‌ها در شبکه ریزتر سطح l+1 را شامل می‌شوند. از این‌رو تعداد تطبیق‌های جدیدی که در این سطح پیداشده است برابر $I_l - I_{l+1}$ است. در نهایت تطابق در همه سطوح با یکدیگر به صورت وزن‌دار جمع می‌شوند تا هسته نهایی تشکیل شود. وزن هر سطح به صورت $\frac{1}{2^{L-l}}$ در نظر گرفته می‌شود و عدم تطابق در سطوح ریزتر مهم‌تر هستند. بنابراین شباهت نهایی میان دو مجموعه X و Y با رابطه (۲-۳) تعریف می‌شود.

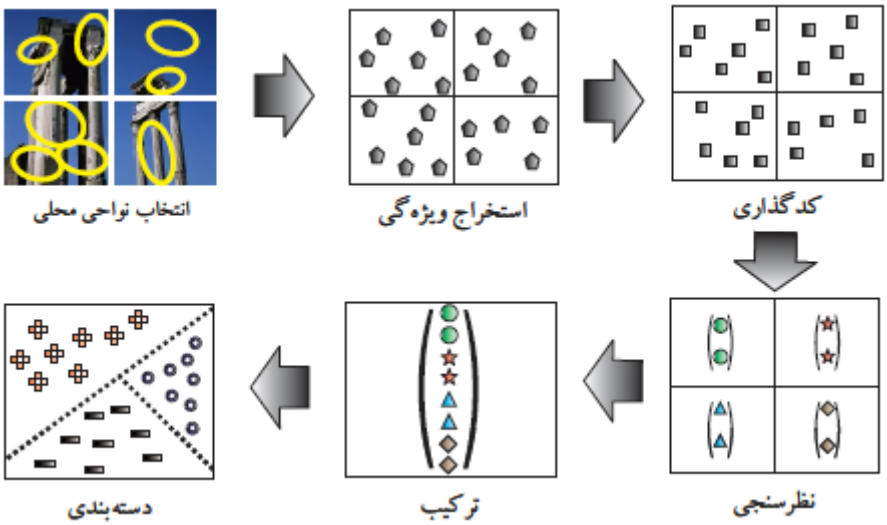
[۱] Spatial Pyramid Matching (SPM)

$$K^L(X,Y) = I^L + \sum_{l=0}^{L-1} \frac{1}{2^{L-l}} (I_l - I_{l+1}) \qquad (۳-۲)$$

همچنین اثبات شده است که این هسته شرط مرسر[1]را دارد [۹].

۲) هرم تطابق مکانی

برای در نظر گرفتن ترتیب مکانی کدها در انتخاب، از روش هسته تطبیق هرمی استفاده شده است. در این روش، شبکه‌بندی بر روی تصاویر انجام می‌شود و ویژگی‌های تطبیقی با تابع f، همان کدهای تصویر هستند. به‌طور دقیق‌تر، اگر X مجموعه ویژگی‌های یک تصویر فرض شود. ویژگی این تصویر در سطح lام با عمل هیستوگرام کدها در هر سلول به دست می‌آید. بنابراین به ازای هر سلول یک بردار ویژگی وجود دارد. در نهایت با پشت سرهم قرار دادن ویژگی‌های همه سلول‌ها در همه سطوح، ویژگی نهایی بدست می‌آید. شکل ۲-۳ کلیه مراحل استخراج ویژگی، با فرض L=1 را نشان می‌دهد. همان‌طور که در شکل ۲-۳ مشخص شده، این روش تنها در مرحله انتخاب با روش مجموعه کلمات تفاوت دارد



شکل ۲-۳ مراحل کلی استخراج ویژگی در روش هرم تطابق مکانی

و بردار ویژگی را به نحوی تولید کرده که مکان ویژگی‌ها نیز در بردار ویژگی نهایی تأثیرگذار باشند. در پایان با استفاده از ماشین بردار پشتیبان و با هسته غیرخطی، مسئله چند دسته‌ای حل و برچسب تصاویر مشخص می‌شوند.

[1] Mercer's condition

۲ - ۲ - ۲روش‌های کدگذاری

در این بخش کدگذاری در مجموعه بسته کلمات توضیح داده می‌شود.

۱) کدگذاری تنک

روش کدگذاری تنک [۷] با تغییر مکانیزم کدگذاری اولیه، باعث کاهش خطای بازسازی می‌شود. لازم به ذکر است که خطای بازسازی مشکل اصلی در کدگذاری روش مجموعه کلمات است. کدگذاری به روش مجموعه کلمات را می‌توان با رابطه بهینه (۲-۴) نوشت.

$$\min_{D,A} \sum_m \|s_m - Da_m\|^2 \qquad (۴-۲)$$

در رابطه (۲-۴)، تنها یک عنصر بردار a_m برابر با ۱ و باقی صفر هستند. s_m و a_m به ترتیب ویژگی و کد متناظر با ناحیه محلی m هستند و D و A ماتریس‌هایی هستند که ستون‌های آن‌ها به ترتیب پایه‌های واژه‌نامه و کدها هستند. منظور از عملگر $\|.\|$نرم دو است. درواقع برای هر m، المانی از a_m برابر یک می‌باشد که پایه متناظر آن در واژه‌نامه D به s_m نزدیکتر است. شرط "تنها یک‌پایه انتخاب شود" شرط محکمی است که می‌توان آن را با شرط "انتخاب تعداد پایه‌ها محدود باشند" متعادل نمود. با انجام این فرایند، یک ویژگی با ترکیب خطی تعداد محدودی پایه ساخته خواهد می‌شود که خطای بازسازی را تا حد زیادی کاهش می‌دهد. این راه‌حل را می‌توان با رابطه بهینه‌سازی (۲-۵) نوشت:

$$\min_{D,A} \sum_m \|s_m - Da_m\|^2 + \lambda|a_m| \qquad \forall_i \|d_i\| \le 1 \qquad (۵-۲)$$

در رابطه (۲-۵) منظور از عملگر $|.|$، نرم یک است و غیرصفر بودن عناصر کمی از بردار a_m را تضمین می‌کند. شرط در نظر گرفته شده بر روی اندازه پایه‌ها، باعث حذف جواب‌های بدیهی می‌شود. با محاسبه کد داده‌ها بر اساس کدگذاری تنک، دقت دسته‌بندی در حد چشم‌گیری افزایش می‌یابد و همچنین دسته‌بندی با یک ماشین بردار پشتیبان با هسته خطی نتایجی مانند هسته غیرخطی دارد. آموزش یک ماشین بردار پشتیبان با هسته غیرخطی هزینه زمانی از مرتبه $O(n^3)$ و آزمون آن هزینه زمانی از مرتبه $O(n)$ دارد، درحالی‌که با هسته خطی هزینه زمانی آموزش به $O(n)$ و آزمون به $O(1)$ کاهش می‌یابد.

ادعا می‌شود که این ویژگی‌ها تا حد زیادی معنایی هستند و به همین دلیل، جداسازی آن‌ها با خط امکان‌پذیر است.

کار دیگر انجام‌شده در این روش، استفاده از عملگر بیشینه‌گیری به جای هیستوگرام در محله انتخاب است. پشتوانه این کار، شبیه بودن عملکرد نورون‌ها در مغز به عمل بیشینه‌گیر است. نتایج عملی نیز پیشرفت محسوس در نتایج دسته‌بندی را نشان می‌دهد.

۲) کدگذاری با پایه‌های محلی

طبق مشاهدات عملی در مرجع [۱۰]، استفاده از پایه‌های محلی برای ساخت یک ویژگی، نتایج دسته-بندی تصویر را نسبت به روش کدگذاری تنک بهبود می‌دهد. از طرف دیگر، روش کدگذاری تنک با وجود دقت دسته‌بندی بالا، از پشتوانه تئوری قوی در یادگیری ماشین برخوردار نیست [۱۰]. در این روش به‌صورت تئوری نشان داده می‌شود که با انتخاب محلی پایه‌هایی برای کدگذاری یک داده و استفاده از یک روش یادگیری خطی، می‌توان هر تابعی را برروی منیفولدی داده یاد گرفت. همچنین پیچیدگی یادگیری، مرتبط با بعد منیفولد است. منظور از منیفولد به هر فضای مجرّد ریاضی اطلاق می‌شود، دارای یک همسایگی (هرچند کوچک) شبیه به فضای اقلیدسی است، ولی از نظر ساختار سراسری می‌تواند از آن پیچیده‌تر باشد. این نتیجه نشان‌دهنده‌ی قرار گرفتن این روش در زمره روش‌های نیمه‌نظارتی است. تابع بهینه‌سازی این روش به‌صورت رابطه (۲-۶) است:

$$\min_{D,A} \sum_m \|s_m - Da_m\|^2 + \lambda \sum_i |a_{mi}| + \|s_m - d_i\|^2 \qquad (۲-۶)$$

در رابطه (۲-۶)، a_{mi} درایه iام بردار a_m است. بنابراین اگر داده s_m از پایه d_i دور باشد، مقدار عبارت $\|s_m - d_i\|$ بزرگ خواهد بود و در بهینه‌سازی با کوچک انتخاب کردن المان متناظر آن، یعنی a_m، مقدار استفاده از این پایه در کدگذاری کاهش می‌یابد. بدین ترتیب در کدگذاری هر داده تنها از پایه‌های محلی آن استفاده می‌شود.

با وجود تئوری دقیق، استفاده عملیاتی از این روش مشکل است. در واقع روش بهینه‌سازی سریعی برای رابطه (۲-۶) وجود ندارد. در نتیجه، مجموعه داده بزرگ با تعداد ویژگی‌های محلی در حدود چند ده میلیون، حل آن در زمان معقول ممکن نیست. در ادامه راهکاری برای حل این مشکل ارائه شده است. ابتدا فرض می‌کنیم که واژه‌نامه مشخص و ثابت است. بنابراین a_m با حل رابطه (۲-۷) به دست می‌آید:

$$\min_{a_m} \sum_m \|s_m - Da_m\|^2 + \lambda \sum_i |a_{mi}| + \|s_m - d_i\|^2 \qquad (۲-۷)$$

ماتریس قطری M را به‌گونه‌ای ایجاد می‌کنیم که iامین عنصر قطر آن برابر با $M_{ii} = \|s_m - d_i\|^2$ باشد. به‌راحتی مشخص است که می‌توان عبارت بالا را با استفاده از ماتریس M به‌صورت رابطه (۲-۸) نوشت:

$$\min_{a_m} \sum_m \|s_m - Da_m\|^2 + \lambda |Ma_m| \qquad (۲-۸)$$

حال با قرار دادن $\beta_m = Ma_m$ داریم:

$$\min_{\beta_m} \sum_m \|s_m - DM^{-1}\beta_m\|^2 + \lambda |\beta_m| \qquad (۲-۹)$$

رابطه (۲-۹) فرمت استاندارد مسئله کدگذاری تنک با واژه‌نامه DM^{-1} است که در بخش قبل شرح داده شد. خوشبختانه برای حل مسئله کدگذاری تنک، روش بهینهٔ LASSO/LARS [۱۱] موجود است. بعد از حل مسئله، کد نهایی به‌صورت $a_m = M^{-1}\beta_m$ محاسبه می‌شود. واژه‌نامه به‌صورت برخط و با استفاده از الگوریتم کاهش تصادفی[۱] یاد گرفته می‌شود. این روش، با وجود بهبود سرعت یادگیری، همچنان برای تعداد داده زیاد، غیرعملی است. به همین دلیل یادگیری واژه‌نامه با استفاده از زیرمجموعه‌ای شامل چند ده‌هزار داده انجام می‌شود.

۳) کدگذاری محلی و خطی

در رابط (۲-۶)، فرم بهینه‌سازی روش کدگذاری با پایه‌های محلی شامل دو جمله است. اولویت جمله اول کاهش خطای بازسازی داده است و اولویت جمله دوم، انتخاب پایه‌های محلی است [۱۲]. با استفاده از این ایده، یک راه‌حل دو مرحله‌ای کدگذاری محلی و خطی پیشنهاد می‌شود:

[۱] Stochastic Gradient Descent

۱. برای ویژگی s_m تعداد k نزدیک‌ترین پایه انتخاب می‌شود و در ستون‌های ماتریس D_{km} قرار می‌گیرد.

۲. عبارت $\min_{\alpha_m} \sum_m \|s_m - D_{knn}\alpha_m\|^2$ با فرض واژه‌نامه ثابت برای محاسبه ضرایب متناظر کدنهایی حل می‌شود.

در گام اول تضمین می‌شود که پایه‌ها محلی انتخاب می‌شوند و در گام دوم ضرایب کد به گونه‌ای انتخاب می‌شوند که خطای بازسازی داده کمینه گردد. در واقع فرم بهینه‌سازی پایه‌های محلی که در یک مرحله حل می‌شد به دو بخش تقسیم شده است. با این روند، دقت دسته‌بندی و زمان اجرا به مقدار قابل توجهی کاهش یافته است که به دلیل اجرای عبارت بهینه‌سازی مرحله دوم تنها برروی تعداد محدودی از پایه‌ها است. برای یادگیری واژه‌نامه نیز یک روش برخط مبتنی بر عبارت بهینه‌سازی اولیه پیشنهاد شده است. در عمل، استفاده از یک روش خوشه‌بندی ساده نیز نتایج مشابهی را تولید می‌کند.

با توجه به سرعت بالای این روش نسبت به روش کدگذاری با پایه‌های محلی، استفاده از آن در دسته‌بندی مجموعه تصاویر بزرگ ممکن است. ایراد این روش تعداد ثابت پایه‌های شرکت‌کننده در کدگذاری است. در روش کدگذاری با پایه‌های محلی، تعداد پایه‌های شرکت‌کننده در ساخت یک داده می‌تواند داده به داده تغییر و با ویژگی آن داده تنظیم شوند.

۴) کدگذاری نرم با برش

درروش کدگذاری نرم، درایه متناظر هر پایه در بردار کد برابر مقدار هسته گاوسی مانند رابطه (۲-۱۰) در نظر گرفته می‌شود [۱۳]:

$$a_{ij} = \exp(-\frac{\|s_i - d_j\|^2}{2\sigma}) \qquad (۲-۱۰)$$

این روش برخلاف روش کدگذاری مجموعه کلمات، تعلق به هر پایه را به صورت نرم و با استفاده از هسته گاوسی اندازه‌گیری می‌کند. با وجود پیشرفت قابل ملاحظه، این روش یک ایراد واضح دارد. از آنجایی که فاصله اقلیدسی فقط در فاصله‌های کوتاه معیار مناسبی برای فاصله منیفولدی است مقدار

رابطه (۲-۱۰) تنها برای مقادیر کوچک $\|s_i - d_j\|^2$ مقادیر درستی است. در مرجع [۱۴] مقدار رابطه (۲-۱۰) تنها برای k نزدیکترین پایه به ویژگی سنجیده می‌شود و باقی درایه‌ها برابر صفر گذاشته می‌شوند. با کمی دقت مشاهده می‌شود که در این روش درایه‌های یکسانی با روش کدگذاری محلی و خطی غیر صفر است و تنها تفاوت در دو روش انتخاب مقادیر درایه‌ها است که در این روش از معیار شباهت و در روش کدگذاری محلی و خطی از ضرایب بازسازی استفاده شده است.

۲ - ۲ - ۳ هسته بهینه برای تشخیص شباهت ویژگی‌ها

در این روش ابتدا به مسئله کدگذاری از نگاه طراحی هسته نگریسته می‌شود و بعد تلاش می‌شود یک هسته بهینه برای حل مسئله ارائه گردد. با این کار یک چارچوب ارائه می‌گردد که می‌توان یک هسته بهینه طراحی نمود در حالی که روش‌هایی که معرفی شد تنها یک راه‌حل ارائه می‌دهند و از این نظر قابل مقایسه و بررسی نیستند. این روش که هم‌زمان با روش کدگذاری محلی و خطی معرفی شده است، با ارائه هسته بهینه سعی در کاهش خطای بازسازی در روش کدگذاری مجموعه کلمات دارد. گرچه نگاه این دو روش به مشکل متفاوت است و راه حل‌های نسبتاً متفاوتی نیز ارائه می‌نمایند اما نتایج عملی مشابهی دارند.

همانطور که در روش مجموعه کلمات گفته شد، در روش هیستوگرام‌گیری ویژگی نهایی از روی بردار کدها به شکل رابطه (۲-۱۱) محاسبه می‌گردد:

$$v_i = \frac{1}{m}\sum_c a_c \qquad (۲-۱۱)$$

با فرض استفاده از هسته خطی در دسته‌بندی می‌توان به رابطه (۲-۱۲) رسید:

$$K(v_k, v_l) = v_k^T v_l = (\frac{1}{m}\sum_c a_c)^T \times \frac{1}{n}\sum_f a_f = \frac{1}{mn} \times \sum_c \sum_f a_c^T a_f = \frac{1}{mn} \times \sum_c \sum_f K_s(s_c, s_f) \qquad (۲-۱۲)$$

که در آن n و m تعداد ویژگی‌های محلی در دو تصویر است. a_c و a_f بردارهای کد استخراجی از تصاویر می‌باشد. s_c و s_f بردار ویژگی SIFT استخراجی از تصاویر می‌باشند. نتیجه این محاسبات این است

که در حالت هیستوگرام‌گیری می‌توان به جای هسته $a_c^T a_f$ از یک هسته دلخواه بر روی دو ویژگی اولیه استفاده کرد. همانطور که گفته شد مشکلی که در استفاده از هسته وجود دارد زمانی بالای فاز یادگیری است، به همین منظور راه‌حلی پیشنهادی می‌گردد که بتوان به‌طور مستقیم فضای نگاشت هسته را پیدا کرد.

هسته $K_s(s_c, s_f)$ هسته‌ای با بعد محدود خوانده می‌شود. اگر نگاشت متناظر آن $\Phi(.)$ بعد محدود داشته باشد، در این صورت می‌توان به صورت رابطه (۲-۱۳) نوشت:

$$K_s(s_c, s_f) = \Phi(s_c)^T \Phi(s_f) \qquad (۲-۱۳)$$

در این صورت با توجه به رابطه (۲-۱۲) می‌توان هسته اصلی بر روی دو ویژگی را به صورت رابطه (۲-۱۴) تعریف کرد:

$$K(v_k, v_l) = \frac{1}{mn} \times \sum_c \sum_f K_s(s_c, s_f) = \frac{1}{m} \sum_c \Phi(s_c) \times \frac{1}{n} \sum_f \Phi(s_f) = \Phi(v_k)^T \Phi(v_l) \qquad (۲-۱۴)$$

اما در عمل فضای نگاشت بسیاری از هسته‌ها بعد محدود ندارند. برای مثال در این روش از هسته گاوسی به‌صورت رابطه (۲-۱۵) استفاده می‌گردد که فضای نگاشت با بعد محدود ندارد. راه‌حلی که پیشنهاد می‌شود یادگیری نگاشتی با بعد محدود است که هسته آن نزدیک به هسته اصلی باشد. در ادامه این روش بررسی می‌گردد.

$$K_s(s_c, s_f) = \exp(-\frac{\|s_c - s_f\|^2}{\sigma}) \qquad (۲-۱۵)$$

۲ - ۲ - ۴ یادگیری ویژگی با بعد محدود

روش کار در یادگیری ویژگی با بعد محدود به این صورت است که ویژگی با بعد زیاد به یک فضای D بعدی تصویر می‌شود. فرض می‌شود پایه‌های این فضا به‌صورت $\{\Phi(z_i)\}_{i=1}^{D}$ داده شده است. اگر فرض شود ستون‌های ماتریس H این پایه‌ها هستند ضرایب تصویر برروی این پایه‌ها به شکل رابطه (۲-۱۶) نوشته و محاسبه می‌گردد.

$$\bar{p}_x = \arg\min_{p_x} \left\| \Phi(x) - Hp_x \right\|^2 \qquad (۱۶-۲)$$

$\bar{p}_x$ ضرایب تصویر بردار $\Phi(x)$ در فضایی است که ستون‌های H را ایجاد می‌کنند. این عبارت بهینه‌سازی محدب و پاسخی به فرم مجموعه دارد که در رابطه (۲-۱۷) نشان داده شده است، این رابطه با مشتق گیری از رابطه (۲-۱۶) و برابر صفر قرار دادن معادله حاصل محاسبه می‌شود:

$$\bar{p}_x = (H^T H)^{-1} H^T \Phi(x) \qquad (۱۷-۲)$$

بنابراین هسته تعریف‌شده در فضای تصویر شده به‌صورت رابطه (۲-۱۸) تعریف می‌گردد:

$$K(x,y) = (H\bar{p}_x)^T (H\bar{p}_y) = \bar{p}_x^T H^T H \bar{p}_y = \bar{p}_x^T L \bar{p}_y \qquad (۱۸-۲)$$

که $L = H^T H$ یک ماتریس $D \times D$ بعدی است که المان (i,j) آن $K(z_i, z_j)$ در فضای نگاشت اولیه است. بنابراین با توجه به قضیه مرسر فضای نگاشت در فضای تصویر به صورت رابطه (۲-۱۹) نوشته می‌شود:

$$\Phi(x) = G\bar{p}_x \qquad (۱۹-۲)$$

که در آن $G = V^{0.5}$ ماتریس $D \times D$ بعدی است که V و S به ترتیب ماتریس ضرایب و مقادیر ویژه در تجزیه طیفی ماتریس L است. سؤال دیگر این است که پایه‌های مناسب برای ایجاد فضا چگونه پیدا می‌شوند. برای این کار استفاده از روش آنالیز مؤلفه‌های اصلی هسته[۱] پیشنهاد می‌گردد. تا پایه‌ها به نحوی انتخاب شوند که در جهت بیشترین واریانس داده‌ها در فضای نگاشت باشند.

برتری تئوری این روش نسبت به روش‌های دیگر این است که در چارچوب هسته بهینه، کدگذاری را بررسی می‌کند، در واقع این چارچوب به اندازه‌ای کلی است که روش‌های دیگر کدگذاری را نیز می‌توان ذیل آن بررسی کرد و با محاسبه هسته آن‌ها قدرت آن‌ها را با هم مقایسه نمود. اما در حیطه عمل با وجود نتایج مشابه روش کدگذاری محلی و خطی، به زمان اجرای بالاتری برای محاسبه ماتریس G، به علت انجام عمل ماتریس معکوس، نیاز دارد.

[۱] Kernel Principal Component Analyses (KPCA)

در این فصل، روش‌های استخراج مجموعه کلمات معرفی شد. روش‌های ارائه شده اغلب برای حل مشکلات موجود در طراحی اولیه روش مجموعه کلمات طرح شدند. یکی از این مشکلات، مستقل بودن ویژگی‌های نهایی نسبت به تغییرات مکانی نقاط محلی است. برای حل این مشکل، روش هرم تطابق مکانی ارائه شد. این روش به همراه روش‌های کدگذاری مورد استفاده قرار می‌گیرد. روش‌های ارائه شده در کدگذاری همگی حول کاهش خطای بازسازی داده، اشتراک دارند. در این میان روش‌های کدگذاری با پایه‌های محلی و روش هسته بهینه برای تشخیص شباهت ویژگی‌ها، از نظر تئوری مورد توجه قرار گرفته‌اند. همچنین روش کدگذاری تنک و کدگذاری محلی و خطی، به لحاظ کارکرد عملی حائز اهمیت هستند. روش هسته بهینه چارچوبی برای مقایسه روش هسته‌ی آن‌ها ارائه می‌دهد، ولی در مورد هسته بهینه نظر ندارد. بنابراین پیشنهاد هسته‌های بهینه‌تر مسئله‌ای باز است.

در مجموع این روش‌ها بر اساس نواحی محلی استخراج برای استخراج ویژگی‌های معنای سطح بالا مناسب می‌باشند.

۲ - ۳ دسته‌بندی صحنه تصویر

دسته‌بندی صحنه تصویر با توجه به افزایش روزافزون تصاویر دیجیتال و پیچیدگی‌های موجود در این حوزه به یک چالش مهم در دهه اخیر تبدیل شده است. در سال‌های اخیر تحقیقات بسیاری در این زمینه انجام شده است. در ادامه فصل کارهای انجام شده در دسته‌بندی صحنه به تفصیل بیان شده است. در دسته‌بندی صحنه تصویر، محاسبه نواحی معنایی سطح بالا مانند نواحی شی امری مهم بوده است. در همین راستا در ادامه، روش‌های محاسبه نواحی کاندیدا به تفصیل بیان می‌گردد. نقاط قوت و ضعف این روش‌ها معرفی و مقایسه‌ای بین این روش‌ها از لحاظ سرعت و دقت انجام می‌شود. در ادامه، کارهای انجام شده در دسته‌بندی تصویر بر اساس پیچیدگی صحنه معرفی شده است. ابتدا به عنوان یک کاربرد در صحنه‌هایی با پیچیدگی پایین، کارهای انجام شده در دسته‌بندی تصاویر نامتعارف معرفی می‌شود و در گام بعدی کارهای انجام شده در صحنه‌هایی با پیچیدگی بالا معرفی می‌گردد. در

این بخش، کارهای انجام شده بر اساس ویژگی‌های معنایی سطح پایین و سطح بالا به صورت جداگانه معرفی می شود. در نهایت روش‌های موجود بر ا ساس مدل‌های مو ضوعی در ت صویر برای د سته‌بندی صحنه‌های پیچیده ارائه می‌گردد. در صحنه‌های پیچیده به د ست آوردن ویژگی‌های معنایی منا سب امری مهم و لازم می‌باشد. در این رساله بر استخراج اشیاء موجود در تصویر به عنوان ویژگی‌های سطح بالا از محتوای تصویر تاکید شده است در همین راستا کارهای انجام شده برای شناسایی و برچسب‌زنی تصاویر نیز معرفی می‌شود.

۲ - ۴ ناحیه کاندیدا

سیستم‌های تشخیص اشیاء که در سال‌های اخیر پیشنهاد شده‌اند، به الگوریتم‌های ناحیه‌های کاندیدا[۱] برای تخمین مکان اشیاء بسیار وابسته می‌باشند. از مهم‌ترین شبکه‌های تشخیص می‌توان به SPPnet [۱۵] و R-CNN[۲] سریع [۱۶] اشاره نمود که در سال‌های اخیر، مورد توجه بسیاری از محققان و کاربرهای صنعتی قرار گرفته است. در این روش‌ها نمی‌توان سراسر تصویر را به دلیل محدودیت‌های زمانی و حالت‌های مختلف از اشیاء جستجو کرد. در همین راستا، ابتدا باید ناحیه‌های کاندید وجود شی استخراج شوند. پیشرفت‌های انجام شده سبب بهبود زمان اجرای شبکه‌های تشخیص شده است و همچنین نشان داده شد که تعیین نواحی کاندیدا مهم‌ترین تنگنای[۳] محاسباتی این‌گونه روش‌ها می‌باشد. نمونه‌ای از ناحیه‌های کاندیدا جهت بررسی وجود شی مورد نظر در شکل ۲-۴ نشان داده شده است. در شکل ۲-۴ سیستم تشخیص شی، انتخاب نهایی خود را از ناحیه‌های کاندیدا انجام می‌دهد.

در مقاله [۱۷]، یک شبکه پیشنهاد ناحیه کاندیدا[۴] (RPN) معرفی شده است. در این شبکه، ویژگی‌های با پیچش‌های مختلف استخراج شده از سراسر تصویر به اشتراک گذاشته می شود. درنتیجه، زمان

---

[۱] Region proposal

[۲] Region-convolutional neural network

[۳] Bottleneck

[۴] Region Proposal Network (RPN)

مربوط به پیشنهادهای ناحیه را قابل چشم‌پوشی می‌سازد. RPN به طور هم‌زمان مرزهای اشیاء و امتیاز[1]شی بودن[2]را در هر موقعیت از تصویر پیش‌بینی می‌کند. اگرچه در مرجع [۱۸] محاسبات بالا در شبکه‌های عصبی پیچش بر پایه ناحیه به عنوان یک اصل بیان شده است ولی این هزینه‌ها با اشتراک کاندیداهای بدست آمده از پیچش به شدت در حال کاهش است [۱۵, ۱۷]. در آزمایش‌های انجام شده در سال‌های اخیر، R-CNN سریع، نرخی نزدیک به زمان بلادرنگ در استفاده از شبکه‌های عمیق در صورتی‌که زمان محاسبه ناحیه‌های پیشنهادی را در نظر نگیریم بدست آورده است. در واقع در این شبکه زمان مربوط به شناسایی ناچیز می‌باشد.

روش‌های نواحی کاندیدا به طور معمول بر ویژگی‌های ارزان و سریع تکیه می‌کنند. جستجو انتخابی [۱۹] یکی از معروف‌ترین روش‌های پیشنهاد ناحیه است که به صورت حریصانه ابرپیکسل‌هایی[3]را که بر اساس ویژگی‌های سطح پایین (مانند میزان شدت روشنایی و رنگ در فضاهای رنگی متفاوت) طراحی شده ادغام می‌کند. اما این روش در مقایسه با شبکه‌های شناسایی کارآمد مانند R-CNN سریع، کندتر می‌باشد. روش پیشنهاد شده در [۲۰]، اخیراً یک تقابل[4]بین کیفیت کاندیدا و زمان را فراهم کرده است. با این‌وجود مراحل تولید نواحی کاندیدا، بار محاسباتی زیادی (ازلحاظ زمان اجرا) در شبکه‌های تشخیص ایجاد می‌کند.

(ج) (ب) (الف)

شکل ۲-۴ نمونه‌ای از ناحیه کاندید. الف) تصویر اصلی، ب) اشیاء موجود در تصویر، ج) ناحیه‌های کاندید [۲۱]

[1] Score
[2] Objectness
[3] Superpixels
[4] Tradeoff

ابرپیکسل به نواحی اتومیک به‌هم پیوسته و دارای مفهوم واحد در تصویر گفته می‌شود.

باید دقت نمود که معمولاً CNN مبتنی بر ناحیه بر اساس[1]GPU توسعه یافته است درحالی‌که روش‌های ناحیه کاندیدا بر روی CPU پیاده‌سازی می‌شوند. درواقع بسیاری از روش‌های کاندیدا قابلیت موازی‌سازی را به خوبی ندارند به همین دلیل چنین مقایسه‌ای بین روش‌های پیشنهاد ناحیه و شناسایی شیء در زمان اجرا نابرابر است. همان‌طور که اشاره شد، بهبود عملکرد سیستم‌های تشخیص و شناسایی اشیاء وابستگی زیادی به روش‌های ناحیه کاندیدا دارند. در همین راستا، در ادامه کارهای انجام شده در این زمینه معرفی می‌شود.

۲ - ۵ کارهای انجام شده در پیشنهاد ناحیه کاندیدا

همان‌طور که در قسمت قبل اشاره شد استخراج ناحیه‌های کاندیدا تأثیر بسیار زیادی در دقت و سرعت روش‌های تشخیص و شناسایی اشیاء در تصاویر دارند. به‌صورت کلی می‌توان روش‌های پیشنهاد نواحی کاندیدا را به چهار دسته‌ی روش‌های مبتنی بر گروه‌بندی، امتیازدهی به پنجره کاندیدا، جایگزین و مرجع تقسیم نمود. در ادامه این روش‌ها معرفی شده است.

۲ - ۵ - ۱ روش‌های مبتنی بر گروه‌بندی

روش‌های مبتنی بر گروه‌بندی برای استخراج ناحیه‌های کاندیدا در سه گروه بر اساس چگونگی تولید کاندیدا، دسته‌بندی می‌شوند. به طورکلی روش‌های تولید کاندیدا بر پایه گروه‌بندی به گروه‌بندی ابرپیکسل[2](SP)، حل کردن مسائل برش گراف چندگانه[3](GC) با[4]seedsهای متنوع یا به طور مستقیم

---

[1] Graphics Processing Unit (GPU)

[2] SuperPixel (SP)

[3] Multiple graph cut
منظور از برش در گراف، تقسیم رئوس گراف به دو زیرمجموعه ناتهی جدا از هم S و (V/S) می‌باشد

[4] Seeds: Superpixels extracted via energy-driven sampling
منظور از Seeds، ابرپیکسل‌های استخراج شده از طریق یک 10نمونه‌گیری انرژی-محور می‌باشد. در مقاله [ ]

از کانتورهای لبه[1] (EC) تقسیم می‌شوند. از مهم‌ترین روش‌های موجود در این حوزه می‌توان به موارد زیر اشاره نمود:

• جستجو انتخابی [۱۹]: در این روش از ادغام ابرپیکسل‌ها، برای تولید کاندیداها استفاده می‌شود. این روش، پارامترهای یادگیری ندارد و از ویژگی‌ها و تابع‌های شباهت برای ادغام ابرپیسکل‌ها استفاده می‌کند. در سال‌های اخیر این روش، به صورت گسترده به عنوان روش کاندیدا برای بسیاری از آشکارسازهای شی مورداستفاده قرارگرفته است.

در مقاله [۱۹] جستجوی انتخابی با ترکیب استراتژی جستجو جامع[2] و قطعه‌بندی ارائه شد. مانند قطعه‌بندی، برای نمونه‌گیری نیز از ساختار تصویر استفاده شد. همچنین از جستجو جامع برای محدود کردن تمامی موقعیت‌های ممکن از اشیاء استفاده شد. نتایج به دست آمده از این روش، بر روی مجموعه کوچکی که وابسته به داده و مستقل از کلاس بوده، بیشترین دقت در مکان‌یابی را نشان داد. آزمایش‌های انجام شده دقتی برابر با ۹۹٪ و متوسط میانگین بهترین همپوشانی بین نواحی پیشنهادی و نواحی صحیح ۰/۸۷۹ در ۱۰۰۹۷ موقعیت را از خود نشان داد. در این روش از یک الگوریتم گروه‌بندی سلسله مراتبی استفاده شد. ناحیه در مقایسه با پیکسل اطلاعات بیشتری دارد، به همین دلیل، در این مقاله از ویژگی‌هایی بر پایه ناحیه بیشتر استفاده شد. روش سریع ارائه شده در مرجع [۲۲] برای انتخاب مجموعه‌ای از ناحیه‌های هم‌پوشان با اشیاء به کار گرفته شد. الگوریتم کلی در شکل ۲-۵ نشان داده شده است. همانطور که اشاره شد، از روش ارائه شده در مرجع [۲۳] برای تولید ناحیه‌های اولیه استفاده شد. سپس از یک الگوریتم حریصانه برای گروه‌بندی ناحیه‌ها با هم استفاده می‌شود.

---

[1] Edge contours — تابع انرژی برپایه شباهت رنگ بین مرزها و هیستوگرام رنگ ابرپیکسل‌ها معرفی شده است.

[2] Exhaustive

Algorithm 1: Hierarchical Grouping Algorithm

Input: (colour) image
Output: Set of object location hypotheses L

Obtain initial regions $R = \{r_1, \cdots, r_n\}$ using [13]
Initialise similarity set $S = \emptyset$
foreach *Neighbouring region pair* (r_i, r_j) **do**
  Calculate similarity $s(r_i, r_j)$
  $S = S \cup s(r_i, r_j)$

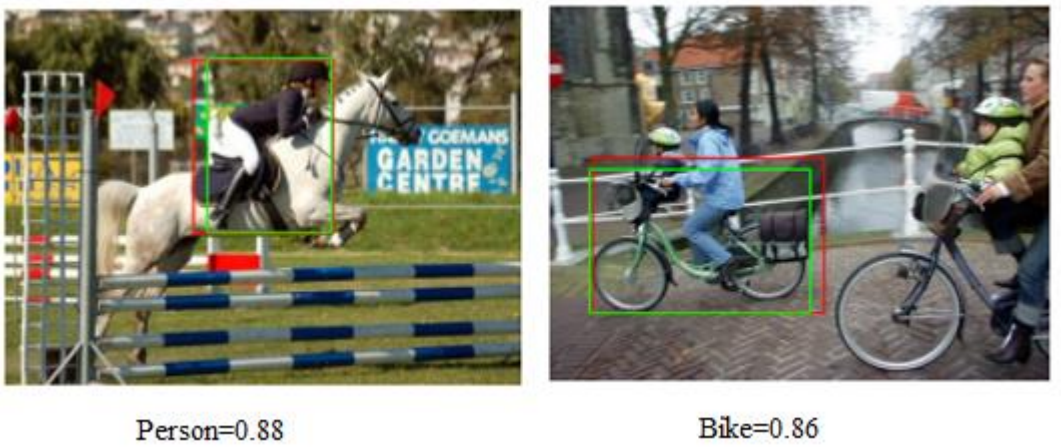
while $S \neq \emptyset$ **do**
  Get highest similarity $s(r_i, r_j) = \max(S)$
  Merge corresponding regions $r_t = r_i \cup r_j$
  Remove similarities regarding $r_i : S = S \setminus s(r_i, r_*)$
  Remove similarities regarding $r_j : S = S \setminus s(r_*, r_j)$
  Calculate similarity set S_t between r_t and its neighbours
  $S = S \cup S_t$
  $R = R \cup r_t$

Extract object location boxes L from all regions in R

شکل ۲-۵ الگوریتم گروه‌بندی سلسله مراتبی در جستجو انتخابی بر اساس مرجع [۱۹]

در این الگوریتم، ابتدا شـــباهت بین همه‌ی ناحیه‌های همســـایه محاســـبه می‌شـــود و در گام بعدی، ناحیه‌هایی با بیشترین شباهت، با هم ترکیب می‌شوند. این کار تا زمانی ادامه می‌یابد که همه‌ی تصویر به یک ناحیه واحد تبدیل شوند. در این مقاله اشاره شد که ناحیه‌های تولید شده می‌توانند در صحنه‌ها و شـرایط روشـنایی متفاوت باشـند. در همین راسـتا الگوریتم گروه‌بندی در فضـاهای رنگی متفاوت با ویژگی‌های مسـتقل و متفاوت اجرا شـد. این ویژگی‌ها شـامل، رنگ در فضـاهای رنگی RGB، شـدت رو شنایی[۱] سطح خاکستری (I) در تصاویر، کانال رنگی rg به همراه I، HSV، RGB نرمال شده و H در HSV می‌با شند. معیار شباهت ا ستفاده شده وابسته به چهار پارامتر رنگ، بافت، اندازه و همپو شانی ناحیه‌ها با هم می‌با شد. شکل ۲-۶ نمونه‌ای از نتایج به د ست آمده از روش ج ستجو انتخابی را ن شان می‌دهد. در هر تصویر محدوده قرمز ناحیه کاندیدا و محدوده سبز مربوط به ناحیه هدف است. مقداری که زیر هر تصویر ذکر شده، میزان همپوشانی این دو محدوده را نشان می‌دهد.

[۱] Intensity



شکل ۲-۶ نمونه‌ای از خروجی جستجو انتخابی در روش گروه‌بندی سلسله مراتبی بر اساس مرجع [۱۹]. محدوده قرمز ناحیه کاندیدا روش جستجو انتخابی و محدوده سبز ناحیه هدف است

• پریم تصادفی:[۱] در مرجع [۲۴] از شباهت ویژگی‌ها مانند روش جستجوی انتخابی استفاده شد. در این روش یک فرایند ادغام جدید به صورت تصادفی برای ابرپیکسل‌ها معرفی شد. در این فرایند ادغام همه‌ی احتمالات ممکن آموزش داده شد. با این روش سرعت به صورت قابل ملاحظه‌ای افزایش پیدا کرد.

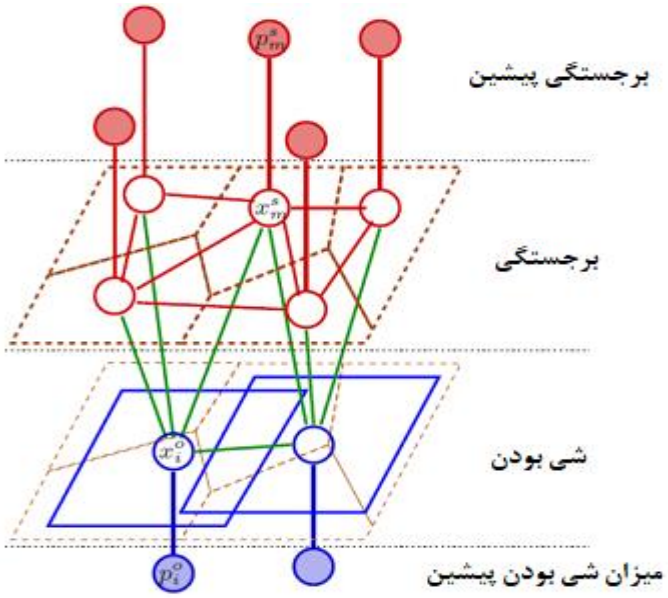
• جستجو محلی و سراسری: در مرجع [۲۵] یک استراتژی مشابه جستجو انتخابی، برای ادغام ابرپیکسل‌ها پیشنهاد شد. ولی از ویژگی‌های مختلف و بیشتری استفاده شده است. در مرحله بعدی، قطعات تولید شده به‌عنوان Seeds ب کار گرفته شد و برای حل کردن برش گراف از CPMC[۲] (در قسمت بعد توضیح داده می‌شود) برای تولید کاندیداهای بیشتر استفاده شد.

• چانگ [۲۶] ترکیبی از برجستگی[۳] و شی‌بودن با مدل گرافیکی، برای ادغام ابرپیکسل در قطعه‌بندی تصویر پس‌زمینه معرفی کرد. مراحل کلی از مدل گرافی پیشنهادی در شکل ۲-۷ نشان داده شده است. این روش با نمونه‌گیری تصادفی از پنجره‌های زیاد شروع می‌شود. به هر پنجره مقدار شی‌بودن و برای هر پیکسل یا ابرپیکسل مقدار برجستگی اختصاص داده می‌شود. برای ارتباط شی‌بودن و برجستگی، مقداری به عنوان سطح برجستگی شیء[۴] برای هر پنجره معرفی شد. این مقدار برای

---

[۱] RandomizedPrim
[۲] Constrained Parametric Min-Cuts (CPMC)
[۳] Saliency
[۴] Object-level saliency

نمایش برجستگی جسم زمینه در هر پنجره استفاده می شود. از شی‌بودن برای تخمین برجستگی استفاده شد. همچنین از برجستگی برای تخمین شی‌بودن استفاده شد. مقدار شی‌بودن هر پنجره در صورتی بالا خواهد بود که برجستگی سطح شیء به خوبی بتواند بسیاری از مقادیر برجستگی پیکسل‌های را نمایش دهد. در مرجع [۲۶]، برجستگی سطح شی توسط اطلاعات به دست آمده از ابرپیکسل‌ها با یک معیار اندازه‌گیری جدید محاسبه شد.



شکل ۲-۷ مراحل کلی الگوریتم چانگ شامل نمونه‌گیری تصادفی جدید و انتصاب مقدار برجستگی به هر پنجره، تخمین شی‌بودن [۲۶]

• قطعه‌بندی خودکار با برش های کمینه[۱](CPMC) [۲۷] از قطعه‌بندی اولیه اجتناب می‌کند و برش‌های گراف را با انواع Seed مستقیماً بر روی پیکسل محاسبه می‌کند. نتایج قطعه‌بندی با حجم بزرگی از ویژگی‌ها رتبه‌بندی می‌شوند.

در مرجع [۲۷] یک روش جدید برای تولید و رتبه‌بندی اشیاء قابل‌قبول در تصویر با فرآیند پایین به بالا[۲] و نشانه‌های سطح متوسط ارائه شد. منظور از نشانه‌های سطح متوسط، میزان شباهت بافت، روشنایی، انرژی کانتور بین ناحیه‌ها و درون ناحیه‌ها می‌باشد.

[۱] Cpmc: Automatic object segmentation using constrained parametric min-cuts

[۲] Down-Top

اشیاء توسط قطعه‌بندی از تصویر استخراج شدند. در این روش قطعه‌بندی به صورت خودکار و بدون دانش اولیه از خواص کلاس اشـــیاء بوده اســـت. برای این منظور از حل دنباله‌ای از محدوده پارامتری برش کمینه بر روی شـــبکه‌ی منظمی[1] از تصـــویر اســـتفاده شـــد. در گام بعدی، برای پیش‌بینی میزان قابل‌قبول بودن قطعه‌ی انتخابی، یک مدل پیوســـته برای رتبه‌بندی اشـــیاء آموزش داده می‌شـــود. این روش برای قطعه‌بندی ســـطح پایین در مجموعه داده VOC09 عملکرد خوبی را از خود نشـــان داد. این روش متوســـط بهترین همپوشـــانی ۰/۷۸ با ۱۵۷ قطعه را دارد. مجموعه داده VOC یک مجموعه داده استاندارد برای شناسایی اشیاء شامل ۲۰ شی متفاوت می‌باشد [۲۸].

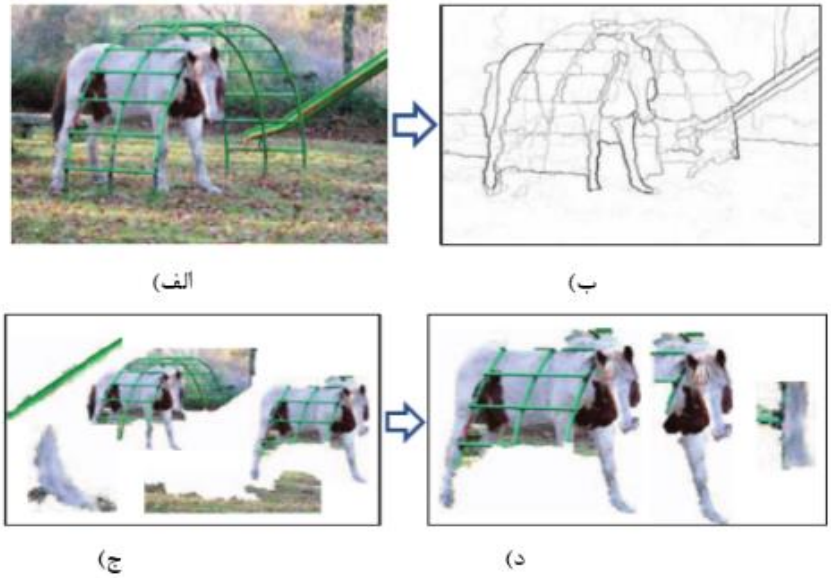
در CMPC برای هر تصـــویر، مجموعه‌ای از پیکســـل‌ها متعلق به پیش‌زمینه فرض می‌شـــوند. این پیش‌زمینه به‌عنوان Seed معرفی می‌شـــود. ســـپس برای هر مجموعه، چندین ســـطح از پیش‌زمینه استخراج می‌شود. هزینه‌های متفاوتی به همه‌ی پیکسل‌های باقیمانده اختصاص داده می‌شود. اما بعضی از پیکسل‌ها در امتداد مرز تصویر، جز Seed منفی می‌شوند.

شکل ۲-۸ مراحل روش قطعه‌بندی خودکار با برش‌های کمینه بر اساس مرجع [۲۷]

از الگوریتم‌ها بر پایه گراف به منظور قطعه‌بندی این Seedها استفاده شد. همچنین میزان شباهت بین پیکسل‌های همسایه به عنوان وزن گراف در نظر گرفته شد. در ادامه یک تابع انرژی تعریف می‌شود. با

[1] grid

این تعریف، برچسب‌های اختصاص داده شده به پیکسل‌ها کاهش داده می شوند و اشیاء مورد نظر استخراج می‌شوند. شکل ۲-۸ روال کلی این روش را نشان می‌دهد.



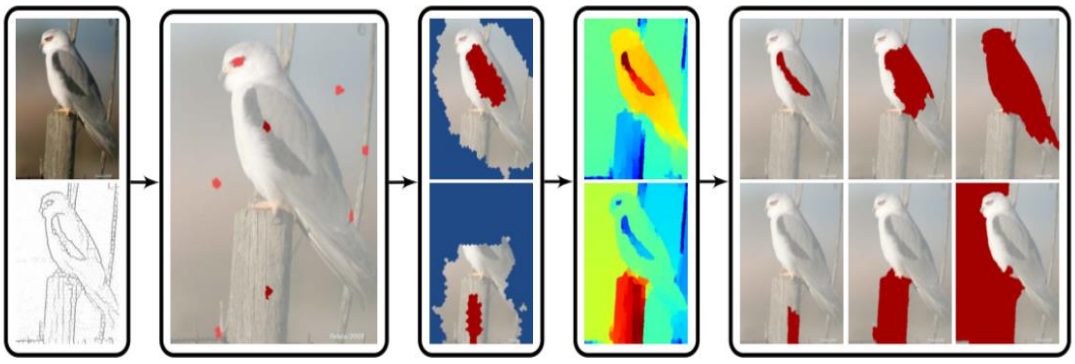
شکل ۲-۹ مراحل مختلف روش قطعه‌بندی سلسله مراتبی از مرزهای انسداد بر اساس مرجع [۲۹]، الف) تصویر ورودی، ب) قطعه‌بندی سلسله‌مراتبی، ج) ناحیه‌های کاندیدا، د) رتبه‌بندی ناحیه‌ها

• در مرجع [۲۹] یک قطعه‌بندی سلسله مراتبی از مرزهای انسداد[۱] و حل برش گراف با انواع Seed و پارامترهای متفاوت، برای تولید قطعات استفاده شد. کاندیداها بر اساس طیف گسترده‌ای از نشانه‌ها رتبه‌بندی شد. این عمل در واقع سبب تشویق تنوع انتخابات می شود. مراحل کلی پیشنهاد شده در مقاله [۲۹] در شکل ۲-۹ نشان داده شد. در این روش، تصویر ورودی بر اساس روش سلسله مراتبی ارائه شده در مرجع [۳۰] قطعه‌بندی می‌شوند. این قطعه تولید شده، ناحیه‌های Seed تصویر را به وجود می‌آورد. در گام بعدی، با استفاده از این Seedها و ویژگی‌های مانند مرز، رنگ و بافت، مجموعه متنوعی از نواحی تولید می‌شوند. این نواحی، می‌توانند به عنوان قطعات شی معرفی شوند که متعلق به اشیاء مختلف هستند. نتایج بدست آمده بر روی مجموعه‌داده BSDS[۲] و PASCAL VOC 2011 [۲۸] نشان دهنده توانایی این روش برای پیدا کردن اشیاء بیشتر با نواحی با اندازه کوچک در پنجره‌های کاندید است. در نهایت با روش یادگیری ساخت‌یافته، ناحیه‌های بدست آمده رتبه‌بندی می‌شوند که به

[۱] Occlusion
[۲] Berekelet Segementation Data Set

احتمال زیاد مناطق با امتیاز بالاتر مربوط به اشـــیاء مختلف هســـتند. نتایج بدســـت آمده بر روی مجموعه‌داده BSDS و PASCAL VOC 2011 نشان دهنده توانایی این روش برای پیدا کردن اشیاء بیشتر با نواحی با اندازه کوچک در پنجره‌های کاندیدا است. مجموعه داده BSDS یک مجموعه داده استاندارد برای برچسب گذاری و قطعه‌بندی تصاویر طبیعی می‌باشد که توسط دانشگاه بریکلی جمع آوری شده است [۳۱] .

• اســـتحکام داده‌ها[۱][۳۲] ، یک روش بهبود یافته از CPMC اســـت که ســـرعت محاســـبه Seedها را با محاســـبه مجدد در ســـراســـر مســـائل برش گراف چندگانه، به صـــورت قابل‌ملاحظه‌ای کاهش می‌دهد. همچنین از روش آ شکار سازی سریع لبه ا ستفاده شد. این الگوریتم بر روی مجموعه داده PASCAL VOC 2012 مورد ارزیابی قرار گرفته اســـت. نتایج گزارش شـــده در این مقاله نشـــان‌دهنده بهبود روش پیشنهادی نسبت به روش‌های دیگر است. بار محاسباتی این روش، در حدود ۲-۴ ثانیه در CPU برای هر تصویر است.



شکل ۲-۱۰ مراحلی کلی روش Geodesic مطرح شده در مرجع [۳۳] (الف) تصویر ورودی و بیش‌قطعه‌بندی آن به همراه نگاشت احتمال مرز، (ب) مکان Seeds، (پ) ماسک پیش‌زمینه و پس‌زمینه تولید شده برای دو Seeds، (ج) استفاده از تبدیل فاصله Geodesic (SGDB) برای این ماسک‌ها، (د) کاندیداهای پیشنهادی با توجه به محاسبه مجموعه‌های سطوح بحرانی در هر SGDB

[۱] Rigor

• روش[1] Geodesic [۳۳] توسط یک روش قطعه‌بندی تصویر پیشنهادی در مرجع [۳۴] شروع می‌شود. طبقه‌بندها از مکان Seedها برای تبدیل فاصله[2] Geodesic استفاده می‌کنند. در این مقاله، Seedها توسط طبقه‌بند بهینه برای کشف اشیاء، آموزش داده شد.

ایده اصلی در این مقاله، معرفی مجموعه‌های سطوح بحرانی[3] در تبدیل فاصله Geodesic است. تبدیل فاصله Geodesic برای مکان‌های Seeds در تصویر محاسبه شد. تبدیل فاصله، در زمان نزدیک برخط[4] محاسبه می شود. همچنین محاسبه هر تبدیل برای تولید کاندیداها در مقیاس‌های متفاوت استفاده می‌شود. ازاین‌رو، استفاده از خط‌لوله[5] در محاسبات، می‌تواند بسیار مؤثر باشد.

همان‌طور که در شکل ۲-۱۰ نشان داده شد، در این روش ابرپیکسل‌های تصاویر ورودی با قطعه‌بندی محاسبه می‌شوند و نگاشت احتمال مرز[6] به عنوان وزن گراف در نظر گرفته می‌شود. در واقع در این گراف هر نود به یک ابرپیکسل مربوط است و هر یال نشان دهنده اتصال‌های بین ابر پیکسل‌های مجاور است. وزن یال‌ها میزان درست‌نمایی[7] مرز اشیاء با توجه به لبه‌های تصویر است.

فاصله Geodestic بین دو نود، اندازه کوتاه‌ترین مسیر بین نودهای موجود است. تبدیل فاصله Geodestic معیاری برای اندازه‌گیری کوتاه‌ترین مسیر بین مجموعه‌ای از نودها تا هر نود است. این معیار را می‌توان توسط الگوریتم دایجسترا در زمان O(nlogn) محاسبه کرد. در این الگوریتم n تعداد ابرپیکسل‌ها است. در صورت وجود شبکه‌های منظم در تصویر، زمان اجرای روش، نزدیک به زمان برخط است. ولی دامنه استفاده شده در این مقاله به صورت منظم نمی‌باشد برای همین منظور باید از روش دقیق‌تری استفاده نمود.

---

[1] ، اندازه کوتاه‌ترین مسیر Geodesic منظور از فاصله بین نودها می‌باشد.

[2] ، معیاری برای Geodesic منظور از تبدیل فاصله اندازه‌گیری کوتاه‌ترین مسیر بین مجموعه‌ای از نودها تا هر نود

[3] Critical level sets

[4] Online

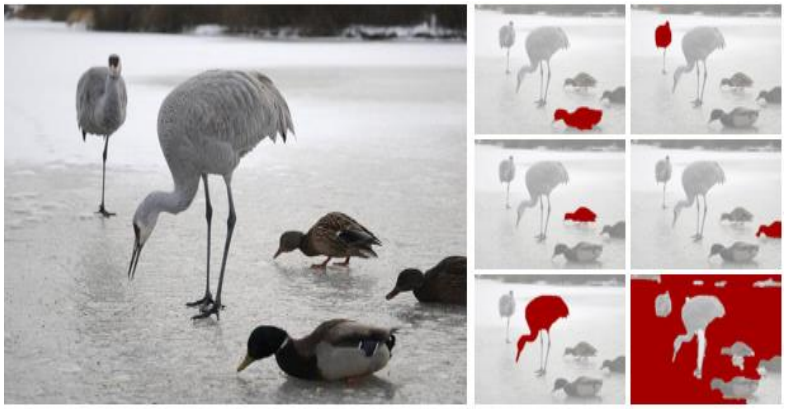
[5] Pipeline

[6] Boundary Probability Map

[7] Likelihood

این تبدیل، به صورت کلی پس‌زمینه و پیش‌زمینه را مشخص می‌کند. هر مجموعه سطح[1] از این تکنیک می‌تواند یک قطعه از تصویر را محدود کند که می‌تواند به عنوان اشیاء کاندیدا در نظر گرفته شوند. منظور از مجموعه سطح، مرز اشیاء موجود در پیش‌زمینه است. در این مقاله پیش‌زمینه و پس‌زمینه بدست آمده از روش قبل با یک روش بر پایه یادگیری بهبود داده شد. در روش پیشنهادی، یک تابع خطی برای پیش‌زمینه و یک تابع خطی دیگر برای پس‌زمینه آموزش داده شد.

نتایج حاصله‌ی روش پیشنهادی، بر روی مجموعه داده PASCAL VOC 2012 در شکل ۲-۱۱ نشان داده شد. آزمایش‌های انجام شده در این مقاله، عملکرد بالا از لحاظ دقت و سرعت را بر روی مجموعه داده PASCAL VOC 2012 گزارش داده است.



شکل ۲-۱۱ اشیاء کاندیدای استخراجی در روش Geodesic بر اساس مرجع [۳۳]. نواحی قرمزاشیاء کاندیدا می‌باشند.

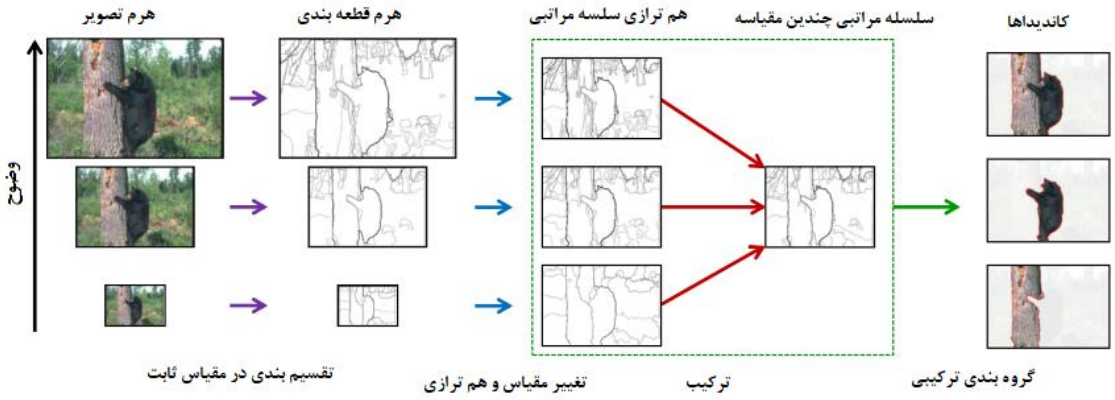
• گروه‌بندی ترکیبی بر پایه چندین مقیاس[2] (MCG) [۳۵] یک الگوریتم سریع برای محاسبه قطعه‌بندی سلسله مراتبی چندین-مقیاسه[3] که در مرجع [۳۴] معرفی شد، ارائه کرده است. قطعات بر اساس استحکام لبه ادغام می شوند. در نتیجه اشیاء بر اساس نشانه‌هایی مانند اندازه، مکان، شکل و استحکام لبه رتبه‌بندی شده‌اند.

[1] Level Set

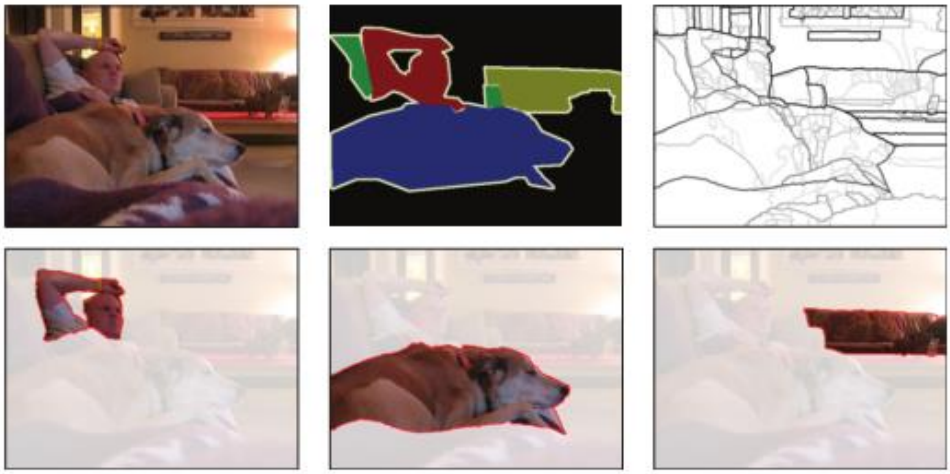
[2] Multiscale Combinatorial Grouping (MCG)

[3] Multi-Scale

مراحل کلی روش پیشنهادی در شکل ۲-۱۲ نشان داده شد. این روش، ابتدا از هرمی در چندین وضوح تصویر شروع می‌شود. در گام بعدی قطعه‌بندی سلسله‌مراتبی در هر مقیاس، به صورت مستقل صورت می‌گیرد. سپس تصاویر حاصل از سلسله مراتبی چندگانه تراز می‌شوند و تصاویر تولیدی در یک تصویر قطعه‌بندی شده با چندین مقیاس تنها ترکیب می شوند. در نهایت با گروه‌بندی اجزای حاصله، لیست امتیازاتی از اشیاء کاندیدا با جستجو در ناحیه‌های فضای ترکیبی محاسبه می شود. این روش بر روی مجموعه داده PASCAL VOC 2012 و BSDS500 مورد ارزیابی قرار گرفته است و عملکرد منا سبی از این روش گزارش شد. نمونه‌ای از نتایج بدست آمده در شکل ۲-۱۳ نشان داده شد.



شکل ۲-۱۲ مراحل کلی روش گروه‌بندی ترکیبی بر پایه چندین مقیاس بر اساس مرجع [۳۵]



شکل ۲-۱۳ نمونه‌ای از خروجی روش گروه‌بندی ترکیبی بر پایه چندین مقیاس بر اساس مرجع [۳۵] سمت بالای تصویر: به ترتیب از سمت چپ: تصویر اصلی، نواحی هدف، قطعه‌بندی سلسله مراتبی در چندین مقیاس، قسمت پایین تصویر: بهترین اشیاء کاندید

۲ - ۵ - ۲روش‌های امتیازدادن به پنجره کاندید

یک روش جایگزین برای تولید کاندیداها، امتیازدهی هر پنجره کاندیدا اســـت. در واقع هر تصـــویر پنجره‌گذاری می‌شـــود و در گام بعدی به هر پنجره امتیازدهی بر اســـاس احتمال شـــی بودن پنجره کاندیدا صورت می‌گیرد. در مقایسه با روش‌های گروه‌بندی، این روش‌ها معمولاً فقط مرز محدوده‌ها را برگشت می‌دهند و سریع‌تر هستند. در مواردی که نمونه‌گیری پنجره با تعداد بالا باشد، این روش‌ها معمولاً پیشنهادهایی با دقت محلی‌سازی پایین تولید می‌کنند. در ادامه به معرفی بعضی از این روش‌ها می‌پردازیم.

• شی‌بودن [۳۶] یکی از اولین و بهترین روش‌های شناخته شده در بحث روش‌های کاندیدا است. یک مجموعه اولیه از کاندیداها از محل‌های برجســـته در تصـــویر انتخاب می‌شـــود. این کاندیداها بر اســـاس چندین نشـــانه مانند رنگ، لبه‌ها، محل، اندازه و اســـتحکام"Superpixel stradding(SS)" امتیازدهی می شوند. برای بررسی استحکام SS معیاری معرفی شده که میزان اشتراک بین ابرپیکسل و پنجره پیشنهادی را محاسبه می‌کند.

• Rahtu [۳۷] با مخزن بزرگی از ناحیه‌های کاندیدا تولید شده توسط ابرپیکسل منحصر به فرد، جفت و سه‌تایی و چندین محدوده انتخاب شده، به صورت تصادفی شروع می شود. استراتژی امتیازدهی بازبینی شده و بهبودهایی در آن ارائه شد.

• Bing [۳۸] از یک طبقه‌بند خطی ساده با ویژگی‌های لبه اســـتفاده می‌کند. در این روش، از پنجره لغزان استفاده شد. سرعت تخمین در این روش بسیار بالا است. در این روش، طبقه‌بند تأثیر اندکی دارد و می‌تواند عملکردی مشـــابه با زمانی که به تصـــویر نگاه نمی‌شـــود را داشـــته باشـــد. این روش‌های مستقل از تصویر، با نام CrackingBing معروف هستند.

• EdgeBoxes [۲۰] همچنین از یک الگو پنجره لغزان بزرگ[۱] شروع می‌شود. اما بر اساس تخمین مرز اشیاء، که توسط ساختار جنگل تصمیم بدست آمده، ساخته شد و یک گام پالایش[۲]بعدی را برای افزایش محلی‌سازی اضافه می‌کند. در این روش، پارامتری یاد گرفته نمی‌شود.

۲ - ۵ - ۳ روش‌های کاندید جایگزین[۳]

• ShapeSharing [۳۹] یک روش غیر پارامتریک[۴]و مبتنی بر داده است که شکل‌های اشیاء را از نمونه‌های درون تصاویر آزمون توسط انطباق لبه‌ها استخراج می‌کند. در نتیجه نواحی در مرحله بعد ادغام می‌شوند و به وسیله حل کردن برش گراف اصلاح می‌شوند.

• Multibox [۴۰] یک شبکه‌عصبی را برای برگرداندن تعداد ثابتی از کاندیداها به صورت مستقیم آموزش داده است. درواقع شبکه بر روی تصویر ورودی لغزانده نشده است. نویسنده بهترین نتایج بر روی دادگان ImageNet را گزارش داده است.

۲ - ۵ - ۴ روش‌های پایه یا مرجع[۵]در پیشنهاد کاندید

بعضی از روش‌ها به عنوان روش پایه و مرجع معرفی می شوند. تمامی این روش‌ها در قسمت‌های قبل مورد ارزیابی قرار گرفته شده‌اند. از روش‌های زیر به‌عنوان روش پایه استفاده می‌شود.

• روش یکنواخت:[۶] در این روش برای تولید کاندیدا، به‌صورت یکنواخت موقعیت مرکز محدوده مرزبندی، ناحیه ریشه مربع و لگاریتم نسبت ابعاد، نمونه‌برداری می شوند. در گام بعدی بازه‌ای از این پارامترها با مجموعه آموزش PASCAL VOC 2007 تخمین زده می‌شوند سپس ۰٫۵٪ از کوچک‌ترین و بزرگ‌ترین مقادیر نادیده گرفته می‌شوند و توزیع تخمین‌زده شده ۹۹٪ از داده‌ها را پوشش می‌دهد.

---

[۱] Coarse
[۲] Refinement
[۳] Alternative Proposal Methods
[۴] Non-parametric
[۵] Baseline
[۶] Uniform

• روش گاوسی[1] به همین ترتیب، یک توزیع گوسین چند متغیره برای موقعیت مرکز محدوده مرزبندی، ناحیه ریشه مربع و لگاریتم نسبت ابعاد تخمین‌زده می شود. در گام بعدی مقدار میانگین و واریانس در مجموعه آموزشی محاسبه می‌شود. در نهایت کاندیداها از این توزیع‌ها نمونه‌برداری شده‌اند.

• روش پنجره لغزان: در این روش پنجره در یک شبکه (گرید) منظم مکان‌یابی می‌شود که برای آشکارسازی شیء در پنجره‌ی لغزان رایج است. تعداد درخواست کاندیداها بر اساس اندازه پنجره توزیع شد. برای هر اندازه پنجره به صورت یکنواخت پنجره‌ها مکان‌یابی شده‌اند. این روش از پیاده‌سازی مرجع [۳۸] الهام گرفته شده است.

• روش ابرپیکسل: همان‌طور که در قسمت‌های قبل بیان شد، ابرپیکسل تأثیر مهمی بر روی رفتار روش‌های پیشنهادی دارند. با توجه به تعریف مرجع [۴۱]، ابرپیکسل به نواحی اتومیک[2] متراکم یا به‌هم پیوسته[3] و ادراکی[4] معنادار[5] در تصویر گفته می‌شود.

۲ - ۶ کنترل تعداد کاندیدهای پیشنهادی

برای مقایسه جامع بین روش‌های مطرح شده باید تعداد کاندیداهای ارائه شده در هر تصویر توسط روش‌هایی کنترل شوند، زیرا برای مقایسه صحیح بین روش‌های موجود باید تعداد کاندیداهای ارائه شده یکسان باشد. در بسیاری از موارد، روش‌ها با ارائه‌ی تعداد کاندیداها بین ۱۰۰ تا ۱۰۰۰۰۰ مورد ارزیابی قرار می‌گیرند. همچنین برخی از روش‌ها امتیاز کاندیداها را ارائه داده و در مقابل بعضی دیگر این توانایی را ندارند. البته نمی‌توان در همه‌ی روش‌ها به صورت صریح و روشن تعداد کاندیداها را کنترل نمود. در روش‌هایی که امتیاز کاندیدا را ارائه می‌دهند، k برترین ناحیه کاندیدا انتخاب می‌شوند. ولی در برخی از روش‌ها، نمی‌توان کنترل مستقیم بر روی تعداد کاندیداها داشت و همچنین امتیاز

---

[1] Gaussiany
[2] Atomic
[3] Compact
[4] Perceptually
[5] Meaningful

کاندیداها نیز مشخص نمی‌شوند. در این روش‌ها می‌توان با کنترل غیرمستقیم به k کاندیدا دسترسی داشت. در این راستا نیاز است که پارامترهای دیگر را تغییر داد تا به این کنترل، دسترسی پیدا نمود. به همین جهت تعداد کاندیداهای تولید شده با پارامترهای متفاوت در زیرمجموعه‌ای از تصویر ذخیره می‌شود. در نهایت با استفاده از یک درون‌یابی خطی بین تنظیم‌های پارامترها، k کاندیدا انتخاب می‌شوند. در بعضی از روش‌ها که هیچ کنترلی بر روی تعداد کاندیدها امکان‌پذیر نمی‌باشد k نمونه به‌صورت تصادفی انتخاب می‌شود.

بعضی از روش‌ها کاندیدهای تکراری تولید می‌کنند که آن‌ها را باید از مجموعه انتخابی در هر روش پاک نمود. مقایسه بین روش‌های ارائه‌شده از لحاظ کنترل‌پذیر بودن تعداد کاندیداها در جدول ۲-۱ نشان داده شد. همچنین در این جدول مقایسه‌ای بین روش‌ها از لحاظ نوع روش و خروجی حاصل صورت گرفته است.

جدول ۲-۱ مقایسه‌ای بین روش‌های استخراج نواحی کاندیدا. علامت -/+ به معنی، دارا بودن یا نبودن خاصیت مشخص است و علامت * در کنترل تعداد کاندید، نشان دهنده‌ی کنترل غیرمستقیم است.

روش	نوع روش	خروجی قطعات	خروجی امتیاز	کنترل
Bing[38]	پنجره لغزان	-	+	+
CPMC[27]	گروه‌بندی	+	+	*
EdgeBox[20]	پنجره لغزان	-	+	+
Endres[29]	گروه‌بندی	+	+	+
Geodesic[33]	گروه‌بندی	+	-	*
MCG[35]	گروه‌بندی	+	+	*
Objectness[36]	پنجره لغزان	-	+	+
Rahtu[37]	پنجره لغزان	-	+	+
RandPrim[24]	گروه‌بندی	+	-	+
Rantalankila[25]	گروه‌بندی	+	-	*
Rigor[32]	گروه‌بندی	+	-	*
SelectiveSearch[19]	گروه‌بندی	+	+	*
Gussian	-	-	-	+
SlidingWindow	-	-	-	+
Superpixels	-	+	-	-
Uniform	-	-	-	+

۲ - ۷ مجموعه‌داده استفاده شده در نواحی کاندیدا

در سال‌های اخیر، تحقیقات و آزمایش‌های بسیاری در زمینه انتخاب ناحیه‌های کاندیدا، تشخیص و شناسایی اشیاء بر روی تصاویر انجام شده است. معمولاً این تحقیقات بر روی مجموعه داده استاندارد بوده که ارزیابی روش های پیشنهادی را با دقت بیشتری امکان پذیر می‌سازد. از مهم‌ترین مجموعه‌داده‌های استفاده شده در سال‌های اخیر می‌توان به ImageNet، PASCAL و MS COCO اشاره نمود. در ادامه به معرفی هر یک از مجموعه‌داده‌ها پرداخته می‌شود و ویژگی هر یک به تفصیل بیان می‌شود.

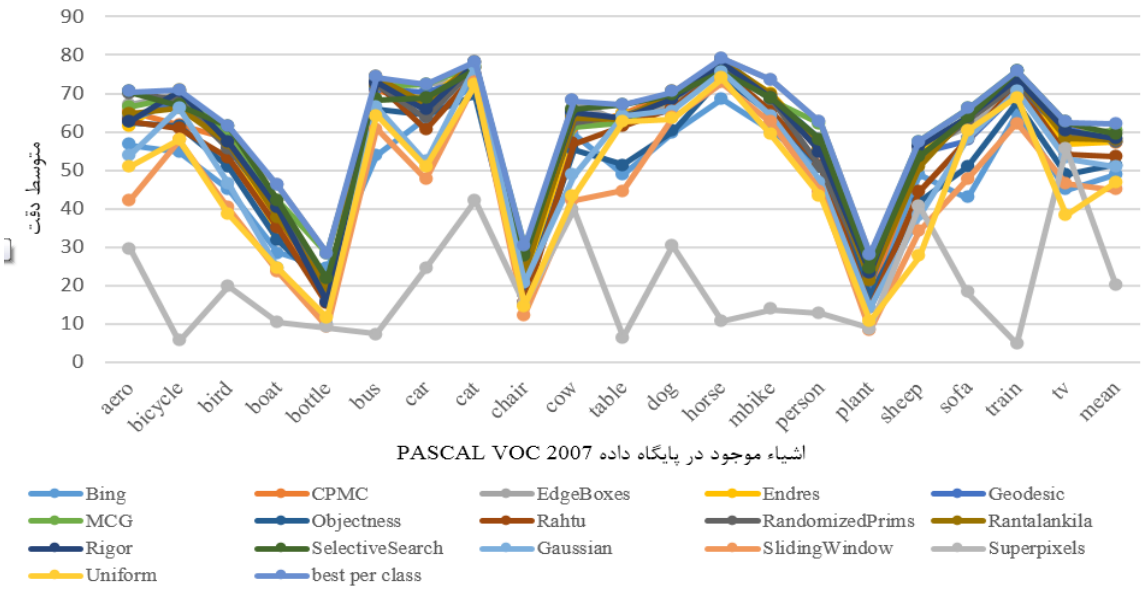
ImageNet: این مجموعه‌داده به عنوان بزرگ‌ترین و گسترده‌ترین مجموعه تصاویر تهیه شده از وب است که برای تشخیص و شناسایی اشیاء مورد استفاده قرار گرفته است. این مجموعه شامل تقریباً ۳,۲ میلیون تصویر در ۵۲۴۷ کلاس متفاوت است. در تحقیقاتی که تاکنون انجام شده تعداد محدودی از این مجموعه داده، مورد استفاده قرار گرفته است. یکی از مجموعه داده‌ای که در این تحقیق مورد بررسی قرار می‌گیرد، مجموعه اعتبارسنجی ImageNet 2013 است که شامل ۲۰۰ کلاس متفاوت و تقریباً بالغبر ۲۰۰۰۰ تصویر نامحدود از این کلاس‌ها است.

PASCAL: این مجموعه داده، شامل ۲۰ کلاس متفاوت است که تقریباً شامل ۵۰۰۰ تصویر نامحدود در حالت‌های مختلف از این کلاس‌ها است.

MS COCO: دو مجموعه داده معرفی شده (ImageNet و PASCAL) برای مقایسه اگر چه از نظر تعداد کلاس‌ها متفاوت بوده ولی از نظر آماری از لحاظ تعداد اشیای موجود در هر تصویر و اندازه اشیاء شبیه هم هستند. مجموعه داده MS COCO شامل ۸۰ کلاس متفاوت از اشیاء است ولی دارای اشیاء بیشتر و اندازه متفاوت‌تر (کوچک‌تر) در هر تصویر نسبت به دو مجموعه داده دیگر است.

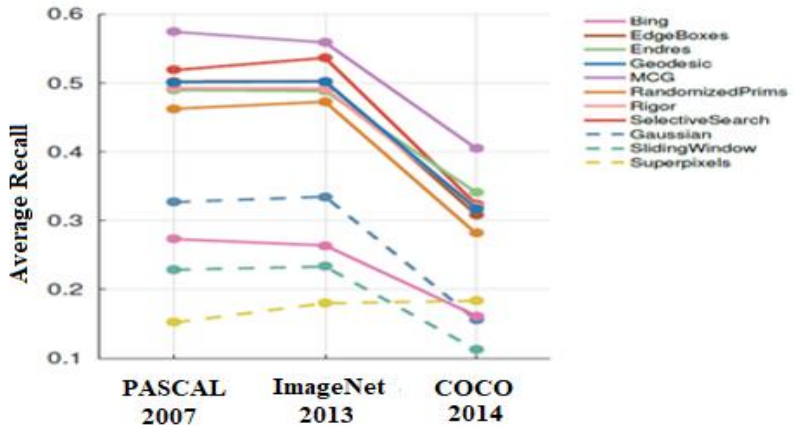
۲ - ۸ مقایسه بین روش‌های نواحی کاندیدا

همان‌طور که در قسمت مقدمه روش‌های نواحی کاندیدا بیان شد، یکی از چالش‌های اساسی و تأثیرگذار در بحث شناسایی اشیاء، پیشنهاد ناحیه‌های کاندیدا بوده است. از طرفی یکی از روش‌های موفق و مشهور مطرح در زمینه شناسایی، روش R-CNN سریع بوده که توسط دانشگاه MIT پیاده سازی شده است. در همین راستا به بررسی تأثیر روش پیشنهاد کاندیدا بر روی این سیستم شناسایی پرداخته شده است. در آزمایش اول از مجموعه داده PASCAL VOC 2007 برای ارزیابی کلاس‌های موجود در این مجموعه داده به صورت جداگانه استفاده شده است. شکل ۲-۱۴ نتایج بدست آمده از روش R-CNN سریع با روش‌های کاندیدای متفاوت را در هر کلاس نشان می‌دهد. با توجه به نتایج بدست آمده روش‌های EdgeBox و MCG در این مجموعه داده بر روی اکثر کلاس‌ها دارای بهترین نتایج بوده و عملکرد مناسبی را از خود نشان داده‌اند. تعداد نواحی کاندیدای استفاده شده در این آزمایش ۲۰۰۰ بوده است. با توجه به نتایج بدست آمده روش R-CNN سریع با نواحی کاندیدا در اشیایی مانند باطری و گیاه نتایج خوبی از خود نشان نداده است ولی در مواردی مانند قطار، اسب، گربه و ماشین این روش نتایج قابل قبولی را از خود نشان داده است.



شکل ۲-۱۴ مقایسه بین روش‌های نواحی کاندیدا بر روی مجموعه داده PASCAL VOC 2007 با روش تشخیص R-CNN سریع

در آزمایش دوم از سه مجموعه داده استفاده شد. شکل ۲-۱۵نتایج بدست آمده بر روی مجموعه PASCAL 2007، ImageNet 2013 و MS COCO را با ۱۰۰۰ ناحیه کاندیدا را نشان می‌دهد. با توجه به نتایج بدست آمده در هر سه مجموعه داده روش MCG بهترین نتایج را از خود نشان داده و روش‌های مرجع و پایه مانند ابرپیکسل، پنجره لغزان و گوسی به تنهایی نتایج قابل قبولی را از خود نشان نمی‌دهند.



شکل ۲-۱۵ نتایج بدست آمده بر روی سه مجموعه داده متفاوت بر اساس ۱۰۰۰ ناحیه کاندیدا و روش‌های متفاوت نواحی کاندیدا

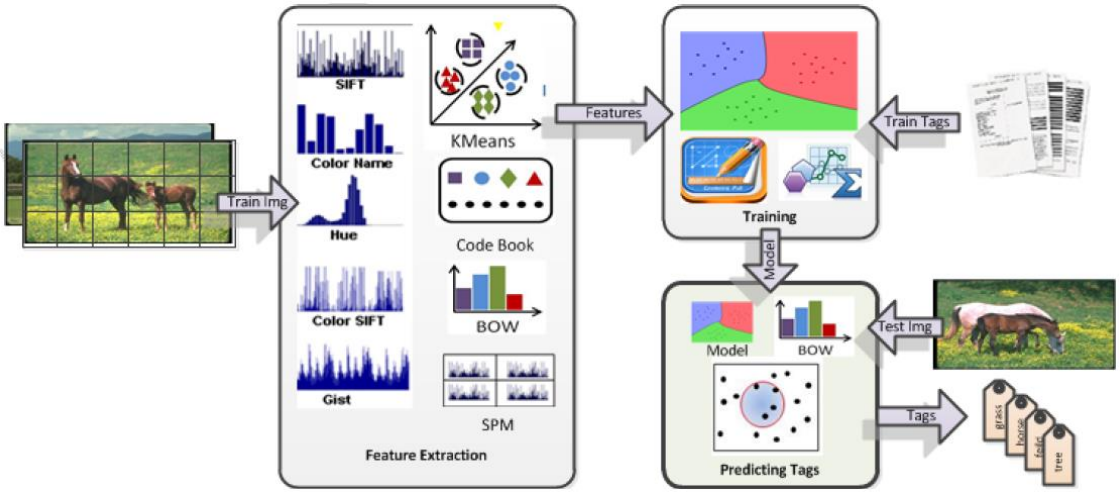
یکی از تنگناه‌های محاسباتی در سیستم‌های تشخیص و شناسایی اشیاء در تصویر، به دست آوردن ناحیه‌های کاندیدا هستند. در واقع عملکرد این نوع سیستم‌ها به روش انتخاب نواحی کاندیدا وابسته است. در همین راستا، به معرفی و مقایسه‌ی روش‌های ناحیه‌های کاندیدا پرداخته و بررسی شد که این روش‌ها در یک مجموعه داده یکسان چه عملکردی را از خود نشان می‌دهند. به منظور مقایسه این روش‌ها، از سیستم شناسایی مشهور و محبوب R-CNN سریع که مورد استقبال بسیاری از محققان و پژوهشگران در این زمینه است، استفاده شد. با توجه به نتایج به دست آمده، استفاده از روش‌های پایه و مرجع به تنهایی عملکرد خوبی نداشته و روش‌های EdgeBoxes و MCG بهترین نتایج را بر روی مجموعه‌داده‌های استاندارد از خود نشان داده‌اند.

۲ - ۹ تشخیص و برچسب زنی اشیاء

هدف از دسته‌بندی تصویر، توصیف سراسری یک تصویر با برچسب توصیف‌کننده است. برچسب‌زنی تصویر، به برچسب‌گذاری محتوایی محلی یک تصویر تاکید دارد. این دو مسئله به هم مرتبط هستند. پس تلاش برای حل مشترک این مسائل امری ضروری است. برای نمونه در یک تصویر با برچسب خیابان، احتمال توصیف تصویر با حضور "ماشین"، "انسان" یا "ساختمان" بیشتر از "ساحل" یا "آب دریا" است. با توجه به توضیحات ارائه شده، بدست آوردن محتوای تصویر در بالا بردن دقت دسته‌بندی بسیار تاثیرگذار است. لازم به ذکر است که منظور از محتوای تصویر، اشیاء موجود در تصویر می‌باشد. در سال‌های اخیر روش‌های زیادی برای تشخیص و برچسب‌زنی اشیاء موجود در تصویر ارائه شده است. برچسب زنی خودکار یکی از کاربردهای یادگیری ماشین محسوب می‌شود و از این لحاظ مانند بسیاری از کاربردهای دیگر می‌توان مراحل کار را به سه گام استخراج ویژگی، آموزش و پیش‌بینی تقسیم نمود [۴۲]. در شکل ۲-۱۶ نمایی از فرایند کلی موجود در طراحی سامانه‌های برچسب‌زنی خودکار نمایش داده شده است. ابتدا ویژگی‌های تصاویر استخراج می‌شوند، سپس بر اساس این ویژگی‌ها و برچسب‌های ثبت شده برای هر تصویر، طی فرآیند آموزش، یک مدل طراحی می‌شود. در مرحله پیش‌بینی برچسب یا آزمایش، ویژگی‌های تصاویر آزمایشی استخراج شده و به این مدل ارائه می‌شود تا برچسب‌هایی برای این تصاویر انتخاب گردد [۴۳]. این مدل ارائه شده بر پایه ویژگی‌های سطح پایین تصویر بوده و دقت روش نسبت به روش‌های نوین عملکرد مناسبی ندارد. در این روش‌ها برای استخراج ویژگی از کل تصویر استفاده شده و برچسب نواحی در مرحله آموزش از قبل آماده بوده است.

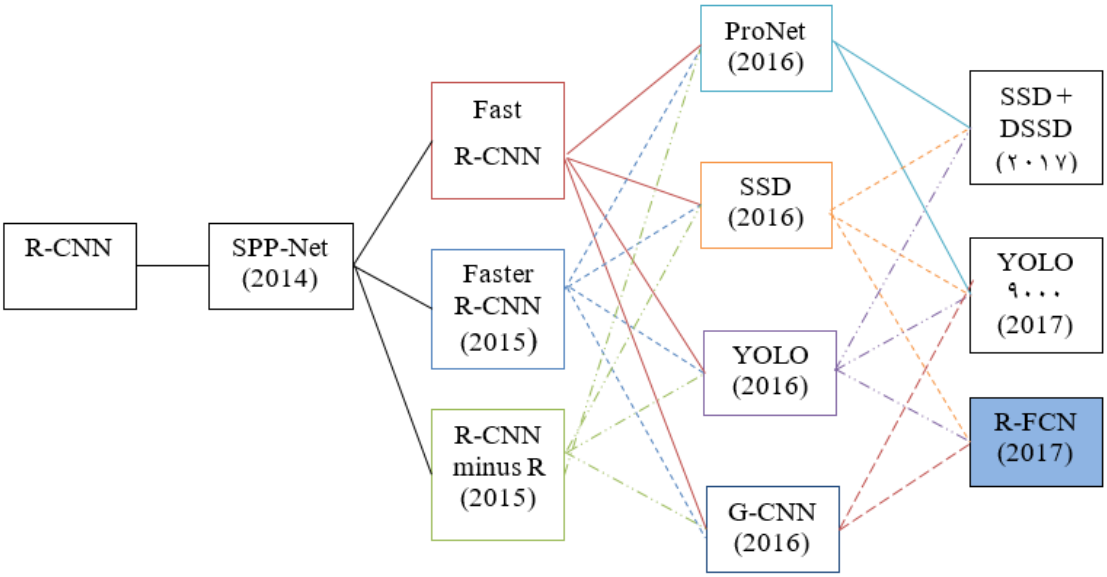
در مرجع [۴۴]، یک روش برپایه اطلاعات متنی یعنی اشیاء دیگر تصویر و صحنه برای تشخیص شی معرفی شده است. در این روش اطلاعات متنی به صورت ماتریس همرخداد مدل شده و هر تصویر توسط برداری که نشان دهنده‌ی اشیاء دیگر تصویر و فرکانس آنها است نشان داده شد. در انتها از روش درخت تصمیم برای تشخیص شی در تصویر استفاده شده است. نتایج این روش بر روی مجموعه داده جمع‌آوری

شده توسط نویسندکان خود مقاله مورد ارزیابی قرار گرفته است. این روش زمانی که تعداد دسته اشیاء زیاد باشد عملکرد خوبی ندارد و روش پیشنهادی وابسته به اطلاعات متنی می‌باشد.



شکل ۲-۱۶ نمایی از فرآیند کلی موجود در سامانه‌های برچسب‌زنی خودکار بر اساس مرجع [۴۳]

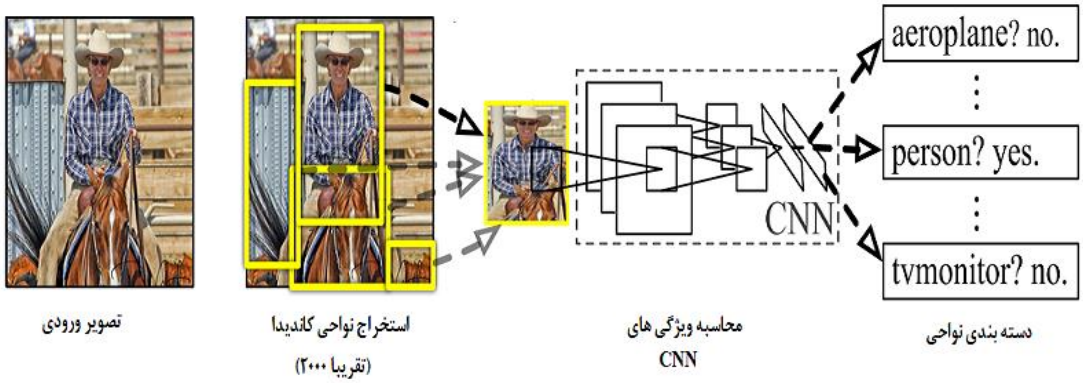
با مطالعات انجام شده در روش‌های موجود به منظور آشکار سازی و برچسب‌زنی اشیاء موجود در تصویر؛ روش‌های پیشنهاد شده بر اساس یادگیری عمیق بهترین عملکرد را از خود در این کاربرد نشان داده‌اند. در همین راستا در ادامه این بخش، کارهای موفق انجام شده در زمینه آشکارسازی و برچسب‌زنی اشیاء بر پایه یادگیر عمیق معرفی می‌شوند.



شکل ۲-۱۷ شمای کلی از کارهای انجام شده برپایه شبکه های عصبی عمیق در آشکارسازی و برچسب‌زنی اشیاء

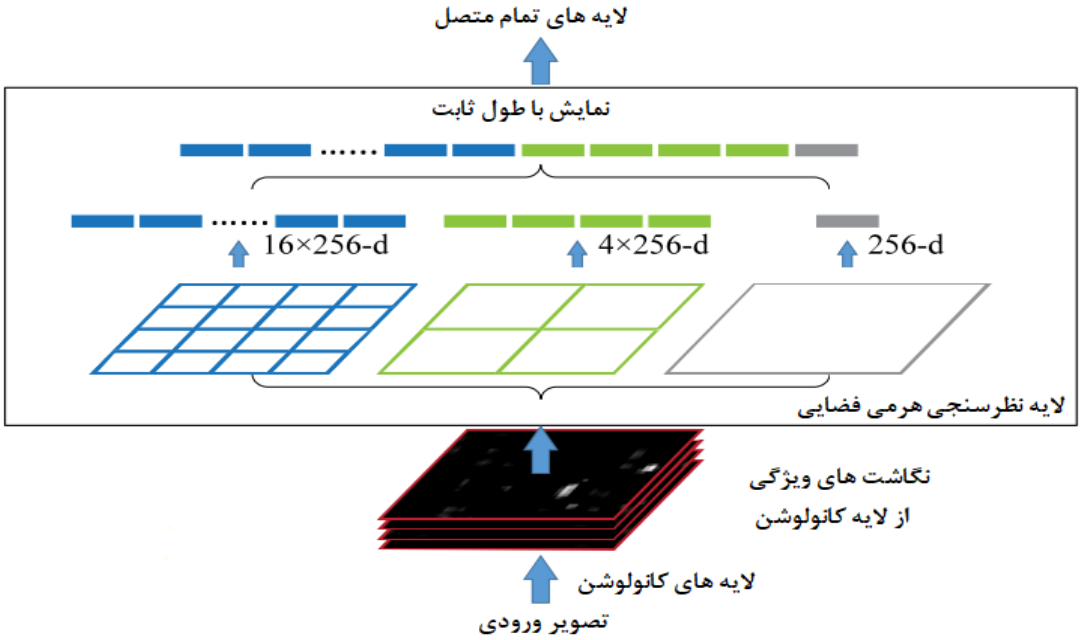
۲ - ۹ - ۱ استفاده از شبکه‌های عمیق در آشکارسازی اشیاء

در سال‌های اخیر استفاده از شبکه‌های عمیق به منظور آشکارسازی و برچسب‌گذاری اشیاء موجود در تصویر مورد توجه محققین و پژوهشگران پردازش تصویر و بینایی ماشین قرار گرفته است. شکل ۲-۱۷ سیر تکامل مقاله‌های منتشر شده در مجلات معتبر دنیا برای شناسایی شی توسط شبکه با مفهوم عمیق را نشان می‌دهد. معماری R-CNN یکی از اولین شبکه‌های عصبی عمیق پیچش بر پایه ناحیه است. در این معماری برای تشخیص اشیاء از شبکه پیچش عمیق با معماری AlexNet استفاده شده است (شکل ۲-۱۸). شبکه R-CNN از نواحی کاندیدا بدست آمده به کمک روش جستجوی انتخابی استفاده می‌کند. در این روش از یک شبکه پیچش برای استخراج ویژگی استفاده می‌شود. همچنین برای دسته‌بندی برچسب اشیاء از یک ماشین بردار پشتیبان خطی باینری استفاده شد.



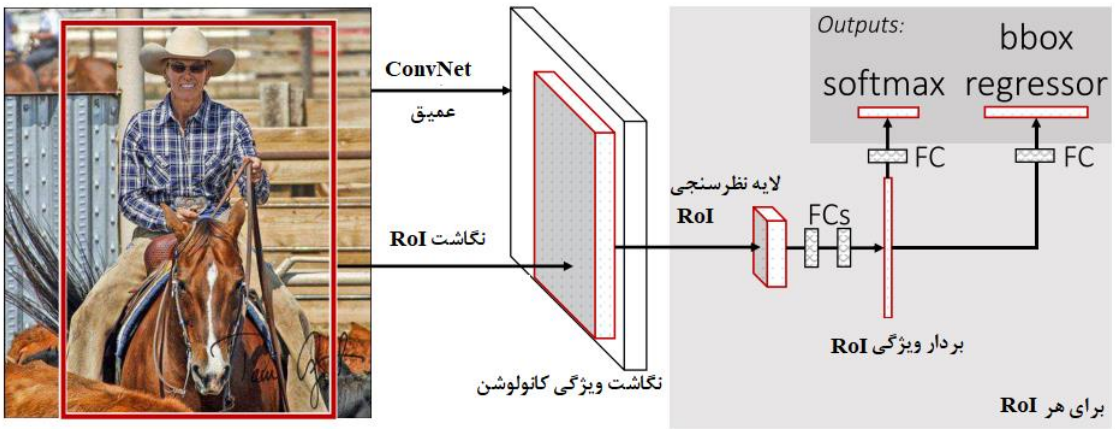
شکل ۲-۱۸ شمای کلی از روش R-CNN ارائه شده در مرجع [۱۸]

شبکه SPP-Net در سال ۲۰۱۴ معرفی شد. این ساختار در تشخیص اشیاء نسبت به شبکه‌های پیشین خود عملکرد بهتری داشت [۱۵]. این روش، ۲۰ تا ۶۰ برابر سریعتر از شبکه R-CNN عمل می‌کند. در این روش، اندازه تصویر ورودی می‌تواند متغیر باشد. در کارهای قبلی، ثابت بودن تصویر ورودی برای لایه تمام متصل امری ضروری بوده است.



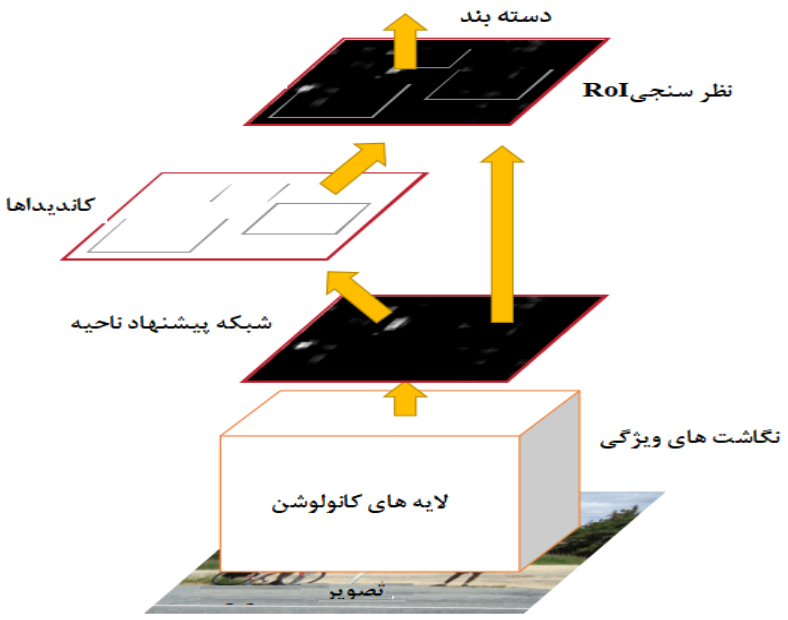
شکل ۲-۱۹ شمای کلی از روش SPP-Net ارائه شده در مرجع [۱۵]

در این روش، با طراحی یک لایه انتخاب هرمی-فضایی قبل از لایه تمام متصل، مشکل ثابت بودن تصویر ورودی رفع شد. در شمای کلی روش مرجع [۱۵] که در شکل ۲-۱۹ نشان داده شد، بعد از لایه پیچش یک لایه هرمی-فضایی تعبیه شد. در شبکه R-CNN به ازای هر ناحیه کاندیدا یک شبکه عمیق آموزش داده می‌شود. در مرجع [۱۵] ادعا شده که تنها یک شبکه به ازای کل تصاویر، آموزش داده می‌شود. سپس از خروجی لایه پیچش، نواحی مربوطه استخراج و برچسب‌گذاری می‌شوند. در واقع در این روش، سرعت بسیار بهبود یافته است .



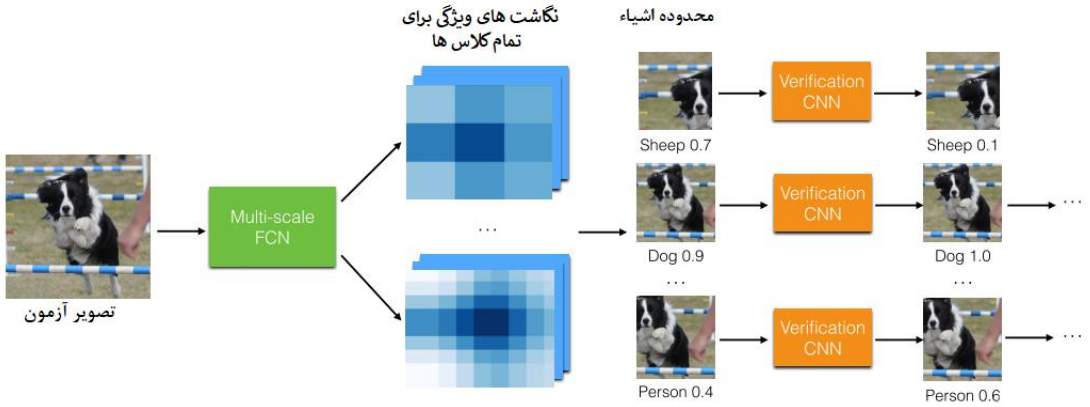
شکل ۲-۲۰ شمای کلی از روش Fast-RCNN ارائه شده در مرجع [۱۶]

با ایده مطرح شده در SPP-Net چندین تحقیق در سال ۲۰۱۵ انجام شد. از معروف‌ترین این شبکه‌ها می‌توان به Fast- RCNN، Faster R-CNN و R-CNN minus R [۴۵] اشاره کرد. معماری مطرح شده در Fast-RCNN در شکل ۲-۲۰ نشان داده شده است. در این معماری، تصویر ورودی و چندین ناحیه مطلوب (RoI) به عنوان ورودی شبکه تمام پیچش هستند. هر RoI توسط لایه انتخاب به یک نقشه ویژگی با طول ثابت تبدیل می‌شود. این نقشه‌ی ویژگی، در گام بعدی توسط لایه‌های تمام متصل به بردار ویژگی نگاشت می‌شود. شبکه برای هر RoI دو بردار خروجی دارد که یک خروجی احتمال هر کلاس و خروجی دیگر نواحی مرتبط به محدوده‌ی شی را مشخص می‌کند. شبکهFaster-RCNN یک شبکه تنها است. در این شبکه بعد از استخراج نقشه‌های ویژگی، شبکه‌ای با عنوان شبکه پیشنهاد ناحیه (RPN) برای استخراج نواحی کاندیدا تعبیه شد. بعد از شبکه پیشنهادی ناحیه لایه انتخاب RoI قرار می‌گیرد. در ادامه مشابه Fast-RCNN موقعیت مرتبط با نواحی محدوده شی تخمین زده می‌شود. در این روش از مفهوم لنگر استفاده شده است. لنگر، هر ناحیه پیشنهاد شده را توسط پنجره لغزان به محدوده‌هایی با اندازه متفاوت از لحاظ نسبت ابعاد تقسیم می‌کند. شکل ۲-۲۱ شمای کلی روش پیشنهاد شده در شبکه Faster-RCNN را نشان می‌دهد.



شکل ۲-۲۱ شمای کلی از روش Faster-RCNN ارائه شده در مرجع [۱۷]

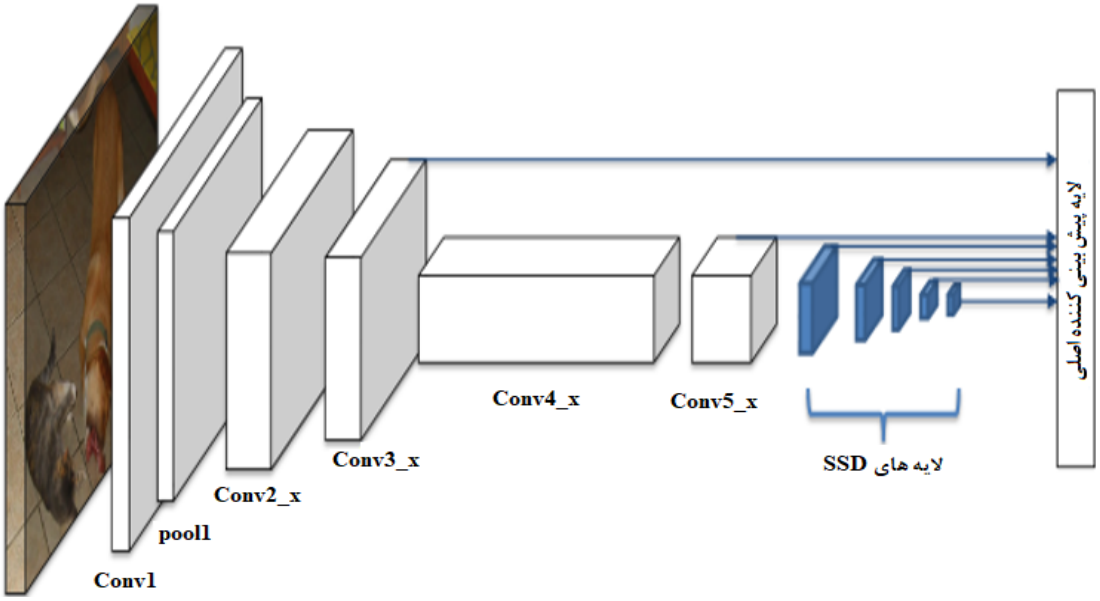
شبکه R-CNN minus R شکل ساده شده‌ی روش RCNN است. در مقاله [۴۵]، نقش تولید نواحی کاندیدا در آشکارسازهای مبتنی بر CNN مورد بررسی قرار گرفته است. در این مقاله، اشاره شد که اطلاعات مهم هندسی در CNN دیده نشده است. در همین راستا، یک آشکارساز جدید برای تولید ناحیه کاندیدا معرفی شد. ترکیب این آشکارساز با SPP-Net عملکرد بهتر و سریعتری ایجاد می‌کند. در ساده‌سازی آشکارسازهای مبتنی بر CNN چندین مرحله یادگیری در یک الگوریتم تلفیق شد. یکی دیگر از شبکه‌های عمیق معرفی شده در سال ۲۰۱۶، ProNet [۴۶] است. در این روش از شبکه‌های عصبی کارامدتر برای پیشنهاد نواحی شامل شی و هم چنین شبکه تمام کانولشن با چندین مقیاس استفاده شد. این روش، از شبکه‌هایی با قدرت بیشتر ولی کندتر برای پیشنهاد ناحیه استفاده می‌کند. این شبکه، امتیازی به محدوده مربوط به اشیاء با در نظر گرفتن مکان‌ها و مقیاس‌های متفاوت اختصاص می‌دهد. برای انتخاب برچسب شی از روش‌های آبشاری یا درختی استفاده می‌شود. در واقع در این روش، نواحی مختلف از یک شی با مکان‌ها و مقیاس‌های متفاوت انتخاب می‌شوند. در گام بعدی با استفاده از CNN احتمال شی بودن و برچسب شی به هر ناحیه اختصاص داده می‌شود. در نهایت برچسب شی و ناحیه مربوط به آن با توجه به ساختار درخت آبشاری انتخاب می‌شود. شمای کلی از این روش در شکل ۲-۲۲ نشان داده شده است.



شکل ۲-۲۲ شمای کلی از روش ProNet ارائه شده در مرجع [۴۶]

شبکه SSD [۴۷] یکی دیگر از شبکه‌های مطرح شده در حوزه تشخیص شی است. این روش تنها از یک شبکه عصبی عمیق برای تشخیص شی استفاده می‌کند. شمای کلی این روش در شکل ۲-۲۳ نشان

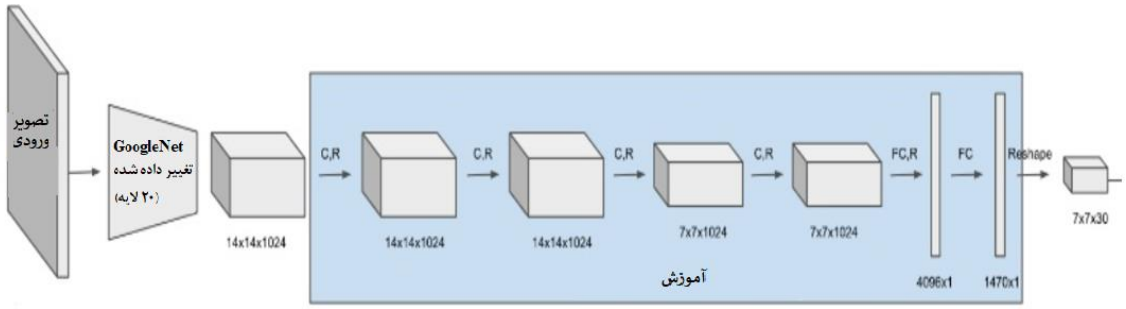
داده شده است. در این روش، یک پیش‌بینی کننده پیچش برای تشخیص شی معرفی شد. در این پیش‌بینی کننده، بالای هر نقشه ویژگی، پیچشی به همراه مجموعه‌ای از فیلترها تعبیه شده است. وظیفه این قسمت، پیش‌بینی دسته‌ی کلاس‌ها و نسبت ابعاد است. در این روش مشابه مفهوم لنگرها در Faster R-CNN استفاده شد. با این تفاوت که SSD این مفهوم را در چندین نقشه ویژگی در تفکیک‌پذیری متفاوت اعمال می‌کند. این روش، عملکرد بهتری از روش Faster R-CNN از خود نشان داد.



شکل ۲-۲۳ شمای کلی از روشSSD ارائه شده در مرجع [۴۷]

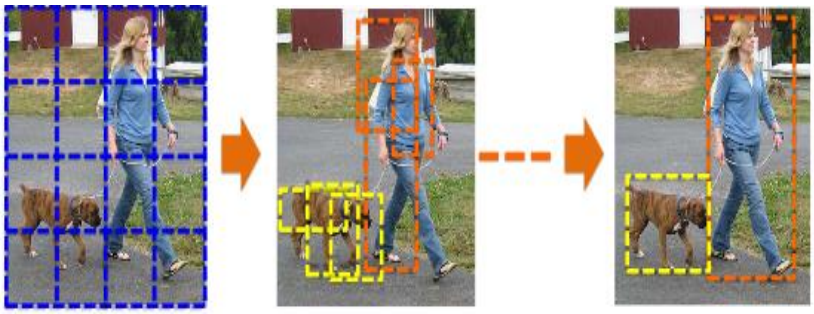
روش YOLO [۴۸] یکی دیگر از روش‌های مطرح شده در سال ۲۰۱۶ است. روش‌های قبلی با دو شبکه پیشنهادی RPN و تخمین کلاس، به دلیل خط لوله کند، سختی در بهینه‌سازی خط لوله‌های مجزا و نیاز به آموزش خط لوله‌ها برای هر بخش، پیچیده بودند. در همین راستا، در این مقاله آشکارسازی به عنوان یک مسئله تخمین در نظر گرفته شد. همچنین شبکه پیچش مجزایی برای طراحی معرفی شد که از لحاظ پیاده‌سازی ساده و سریع است. در این روش، تصویر به تعداد ثابتی شبکه مشبک تبدیل می‌شود. در صورتی که مرکز یک شی در محدوده یک شبکه مشبک قرار بگیرد این شبکه می‌تواند برای تشخیص شی قابل قبول باشد. در ادامه برای هر شبکه مشبک قابل قبول، تعداد ثابتی محدوده در نظر گرفته می‌شود که از معماری شکل ۲-۲۴ برای تشخیص شی بودن استفاده می‌شود. معماری پیشنهادی شامل ۲۴ لایه پیچش و دو لایه تمام متصل می‌باشد، و ۲۰ لایه اول مشابه معماری GoogleNet

می‌باشد. محدودیت این روش در تشخیص اشیاء با اندازه کوچک و اشیاء با ابعاد متفاوت است. همچنین تابع خطا استفاده شده تخمینی بوده و میزان خطای تخمینی برای اشیاء با نسبت ابعاد متفاوت می‌تواند یکسان باشد.



شکل ۲-۲۴ شمای کلی از روش YOLO ارائه شده در مرجع [۴۸]

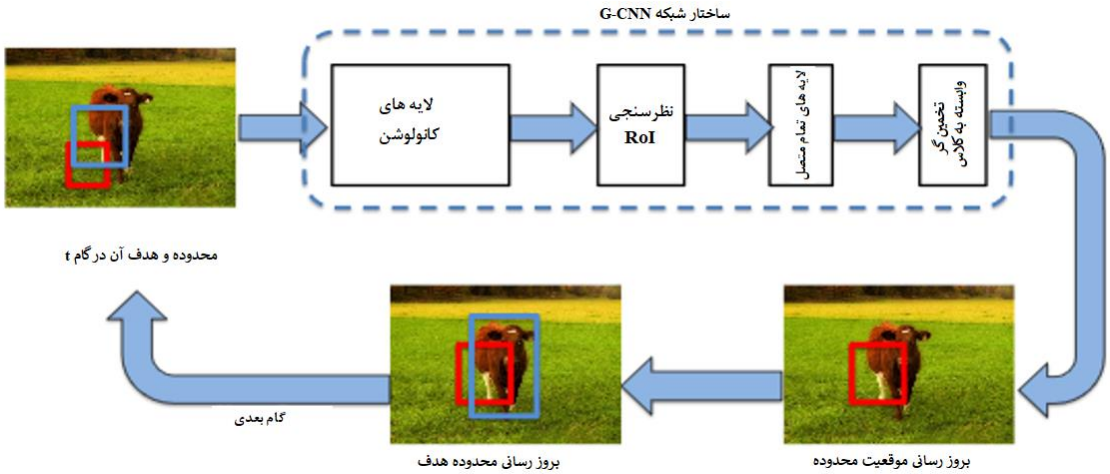
روشG-CNN [۴۹] یک تکنیک تشخیص برپایه CNN و بدون استفاده از الگوریتم پیشنهاد کاندیدا است. همانطور که در شکل ۲-۲۵ نشان داده شد G-CNN با یک شبکه مشبک چند مقیاسه با محدوده ثابت شروع می‌شود. یک رگرسیون تکراری برای جابجایی محدوده و مقیاس عناصر شبکه به منظور استخراج اشیاء آموزش داده می‌شود. در واقع G-CNN مسئله تشخیص اشیاء را به عنوان پیدا کردن مسیر از شبکه مشبک ثابت به محدوده اطراف اشیاء مدل می‌کند .G-CNN با محدوده حدود ۱۸۰ ناحیه در شبکه‌های مشبک با مقیاس متفاوت، عملکردی مشابه با روش Fast R-CNN با ۲۰۰۰ محدوده تولید شده توسط تکنیک‌های پیشنهاد کاندیدا دارد.



شکل ۲-۲۵ مراحل انتخاب نواحی در G-CNN ارائه شده در مرجع [۴۹]

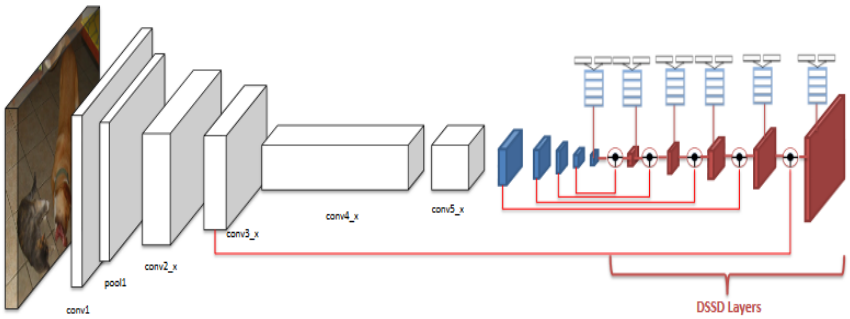
در واقع این روش، بدلیل حذف مرحله پیشنهاد ناحیه و کاهش تعداد نواحی پیشنهادی سریعتر از روش Fast R-CNNاست. شکل ۲-۲۶ شمای کلی این روش برای تشخیص شی را نشان می‌دهد. در این روش یک تابع هدف تعریف می‌شود و با استفاده از شیب نزولی تصادفی بهینه می‌شود. در همین راستا،

G-CNN در مرحله آموزش با تعدادی تکرار، محدوده تخمین را به محدوده واقعی نزدیک می‌کند. معماری شبکه CNN استفاده شده در این روش AlexNet و VGG است .



شکل ۲-۲۶ شمای کلی از روش G-CNN ارائه شده در مرجع [۴۹]

از مهمترین کارهای انجام شده در سال ۲۰۱۷ می‌توان به روشSSD+DSSD [۵۰]، YOLO9000 [۵۱] و RFCN [۵۲] اشاره کرد. در روش DSSD ابتدا دسته‌بند پیشرفته Residual-101 [۵۳] با سیستم تشخیص سریع SSD ترکیب شده است. در ادامه شبکه SSD+ Residual-101 با معرفی لایه‌های پیچش معکوس تقویت شده تا بتواند عملکرد خوبی را در تشخیص اشیاء کوچک و بهبود عملکرد تشخیص شی داشته باشد. روش پیشنهادی در شکل ۲-۲۷ نشان داده شده است .

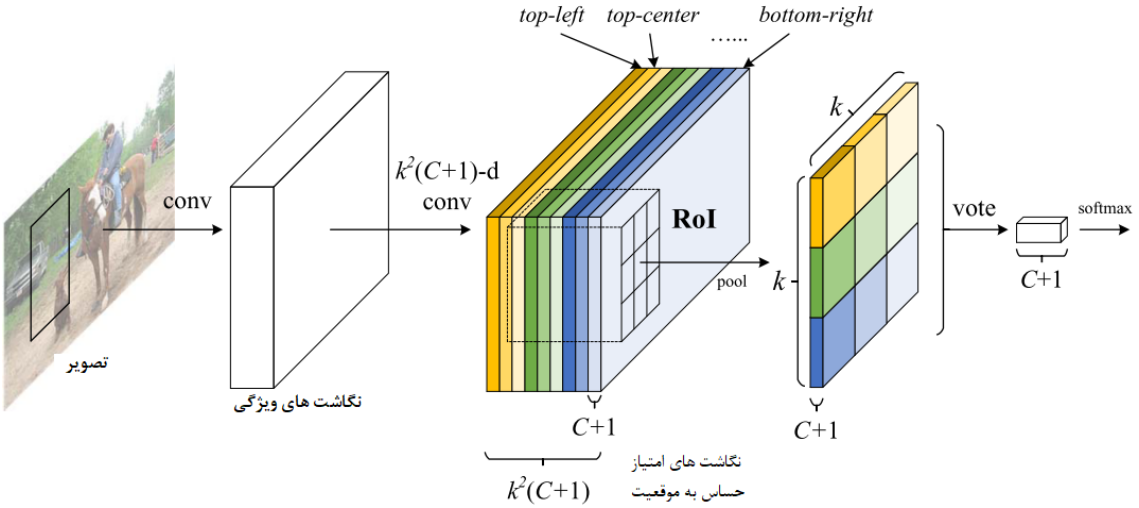


شکل ۲-۲۷ شمای کلی روش SSD+DSSD ارائه شده در مرجع [۵۰]

روش YOLO9000 نسبت به YOLO سریعتر، قویتر و بهتر است. در این روش از نرمال‌سازی دسته‌ای، دسته‌بند با تفکیک‌پذیری بالاتر، پیچش با محدوده‌های لنگر، آموزش با چندین مقیاس، استفاده از دسته‌بند سلسله مراتبی و اتصال دسته‌بندی با آشکارسازی استفاده شد.

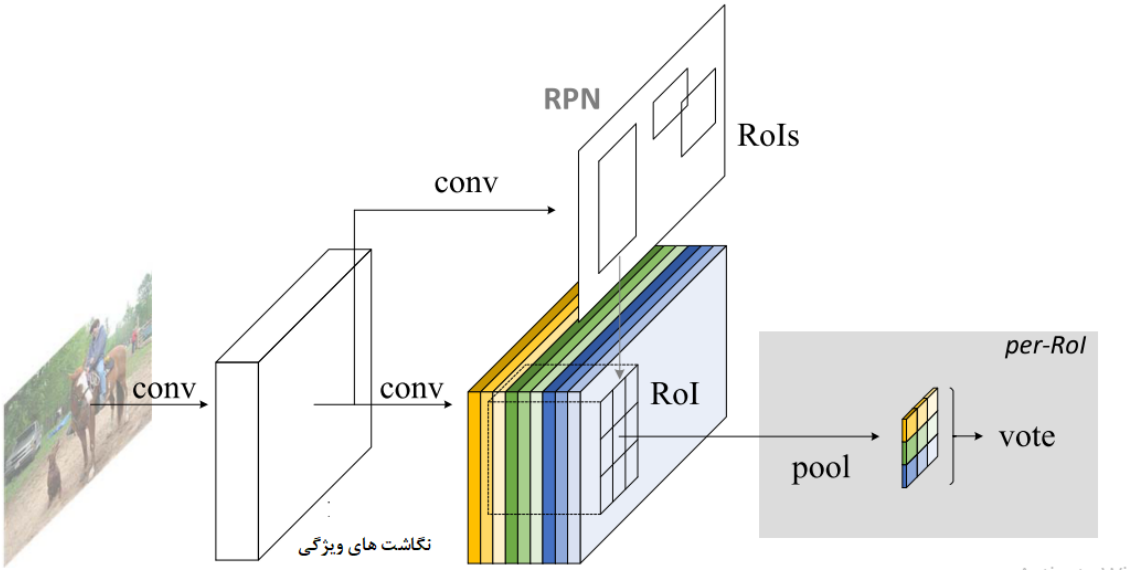
با توجه به مطالب ارائه شده در این بخش، استفاده از شبکه‌های عصبی عمیق برای تشخیص شی را می‌توان به وسیله لایه انتخاب ناحیه مطلوب به دو زیرشبکه تقسیم نمود [۱۸]. دسته اول، زیرشبکه تمام پیچش مستقل از RoIs با اشتراک‌گذاری محاسبات و دسته دوم زیرشبکه بر پایه RoI بدون اشتراک‌گذاری محاسبات است. این تحلیل به لحاظ تاریخی از معماری‌های پیشگام طبقه‌بندی مانند AlexNetو VGG Nets بدست آمد. این شبکه‌ها از دو زیرشبکه تشکیل شده است. یک زیرشبکه پیچش که در انتهای آن یک لایه انتخاب فضایی تعبیه شده و زیرشبکه دیگر شامل چندین لایه تمام متصل است. بنابراین آخرین لایه انتخاب فضایی در شبکه‌های دسته‌بندی تصویر به لایه انتخاب RoI در شبکه‌های تشخیص اشیاء تبدیل می‌شوند [۱۶-۱۸]. اما شبکه‌های دسته‌بندی تصویر ResNets و GoogleNets در طراحی خود از لایه تمام پیچش استفاده می‌کنند.

در این ساختار، فقط آخرین لایه به صورت تمام متصل است. این لایه برای تشخیص اشیاء به شبکه اضافه می‌شود. در این نوع از شبکه‌ها، از لایه‌های تمام پیچش برای ایجاد اشتراک‌گذاری، زیرشبکه پیچش در معماری تشخیص شیء و زیرشبکه مبتنی بر RoI بدون لایه مخفی استفاده می‌شود. با این وجود، در مرجع [۵۲] اشاره شد که اشیایی که با دقت طبقه‌بندی برتر شبکه مطابقت ندارند دارای دقت تشخیص پایین‌ترند. برای اصلاح این مسئله، در مقاله ResNet ، لایه انتخاب RoI آشکارساز Faster R-CNN به صورت غیرطبیعی بین دو مجموعه لایه‌های پیچش قرار داده شد. این کار، زیرشبکه مبتنی بر RoI عمیق‌تری را ایجاد می‌کند که سبب بهبود دقت می‌شود. محاسبات هر RoI اشتراک گذاشته نمی‌شود و سرعت سیستم را کاهش می‌دهد.



شکل ۲-۲۸ شمای کلی از روش نگاشت‌های امتیازدهی حساس به موقعیت ارائه شده در مرجع [۵۲]

یکی دیگر از روش‌های مشهور ارائه شده در کاربرد شناسایی اشیاء روش R-FCN است. این روش در سال ۲۰۱۷ معرفی شد. این روش بهترین عملکرد را نسبت به روش‌های مطرح در سال‌های اخیر داشته است. در روش R-FCN از شبکه‌های پیچش کاملا متصل مبتنی بر ناحیه برای تشخیص دقیق و کارآمد شیء استفاده شد. این روش در مقایسه با آشکارسازهای قبلی مبتنی بر ناحیه مانند Fast / Faster R-CNN، از پیچش‌های کاملا متصل استفاده می‌کند. این پیچش‌های کاملا متصل، تقریبا تمام محاسبات را در تمام تصویر به اشتراک می‌گذارند. برای رسیدن به این هدف، از مفهومی به عنوان نگاشت‌های امتیازدهی حساس به موقعیت استفاده شد. در شکل ۲-۲۸ شمای کلی از این روش نشان داده شده است. در این روش، کانال‌ها، نشان دهنده مکان هر RoI در تصویر ورودی می‌باشند. هر کانال برای مکان خاصی طراحی شد. معماری کلی این روش، در شکل ۲-۲۹ نشان داده شده است. همانطور که در شکل مشخص است، یک شبکه RPN برای پیشنهاد RoIهای کاندیدا معرفی شد. RoIهای کاندیدا به نقشه امتیاز اعمال می‌شوند. در این شبکه وزن لایه‌های قابل یادگیری، پیچش‌ها هستند که برای تصویر ورودی محاسبه می‌شوند. هزینه محاسبه هر RoI ناچیز است. RPN یک شبکه تمام پیچش است. در این شبکه، ویژگی‌های استخراج شده بین RPN و R-FCN به اشتراک گذاشته می‌شود. معماری R-FCN به منظور دسته‌بندی RoIها به دسته‌های اشیاء و پس‌زمینه، طراحی شده است.



شکل ۲-۲۹ شمای کلی از روش R-FCN ارائه شده در مرجع [۵۲].

آخرین لایه پیچش، مجموعه‌ای از ۲۰۰۰ نگاشت امتیازدهی حساس به موقعیت می‌باشد که برای هر دسته (C) ارائه خواهد داد. در واقع تعداد کانال لایه خروجی، $(C+1)k^2$ است. مجموعه‌ای از k نقشه‌های امتیاز به k*k شبکه مشبک فضایی توصیف‌کننده مکان نسبی وابسته است. برای نمونه به ازای k=3، نه نگاشت امتیاز به صورت {top-left, top-center,top-right, ..., bottom-right} برای هر کلاس اختصاص داده می‌شود. آخرین لایه R-FCN، یک لایه انتخاب RoI حساس به موقعیت است. این لایه خروجی آخرین لایه پیچش و تولید کننده امتیاز برای هر RoI را تجمیع می‌کند. شبکه عمیق استفاده شده در این مقاله دارای معماری ResNet-101 است. ResNet-101 شامل ۱۰۰ لایه پیچش، به دنبال آن لایه انتخاب متوسط سراسری و لایه تمام متصل ۱۰۰۰ کلاسه است. جدول ۲-۲ نمایی کلی از ساختار شبکه‌های ResNet را نشان می‌دهد.

جدول ۲-۲ معماری ResNet برای آموزش ImageNet [۵۳]

Layer name	Output size	18-layer	34-layer	50-layer	101-layer
Conv1	112×112	$7\times7,64$ Stride 2			
Conv2.x	56×56	3×3 Max pool Stride 2			
		$\begin{bmatrix}3\times3,64\\3\times3,64\end{bmatrix}\times2$	$\begin{bmatrix}3\times3,64\\3\times3,64\end{bmatrix}\times3$	$\begin{bmatrix}1\times1,64\\3\times3,64\\1\times1,256\end{bmatrix}\times3$	$\begin{bmatrix}1\times1,64\\3\times3,64\\1\times1,256\end{bmatrix}\times3$
Conv3.x	28×28	$\begin{bmatrix}3\times3,128\\3\times3,128\end{bmatrix}\times2$	$\begin{bmatrix}3\times3,128\\3\times3,128\end{bmatrix}\times4$	$\begin{bmatrix}1\times1,128\\3\times3,128\\1\times1,512\end{bmatrix}\times4$	$\begin{bmatrix}1\times1,128\\3\times3,128\\1\times1,512\end{bmatrix}\times4$
Conv4.x	14×14	$\begin{bmatrix}3\times3,256\\3\times3,256\end{bmatrix}\times2$	$\begin{bmatrix}3\times3,256\\3\times3,256\end{bmatrix}\times6$	$\begin{bmatrix}1\times1,256\\3\times3,256\\1\times1,1024\end{bmatrix}\times6$	$\begin{bmatrix}1\times1,256\\3\times3,256\\1\times1,1024\end{bmatrix}\times23$
Conv5.x	7×7	$\begin{bmatrix}3\times3,512\\3\times3,512\end{bmatrix}\times2$	$\begin{bmatrix}3\times3,512\\3\times3,512\end{bmatrix}\times3$	$\begin{bmatrix}1\times1,512\\3\times3,512\\1\times1,2048\end{bmatrix}\times3$	$\begin{bmatrix}1\times1,512\\3\times3,512\\1\times1,2048\end{bmatrix}\times3$
	1×1	Average pool, 1000-d fc, softmax			
FLOPs		$1.8*10^9$	$3.6*10^9$	$3.8*10^9$	$7.6*10^9$

در ادامه این فصل، سیستم‌های دسته‌بندی تصویر بر اساس محتوا معرفی می‌شود. در گام نخست روش-های موجود برای دسته‌بندی صحنه‌هایی با پیچیدگی کم مانند تصاویر دو کلاسه با عنوان تصاویر نامتعارف بررسی می‌شوند. در گام بعدی روش‌ها و تحقیقات انجام شده در سال‌های اخیر در دسته‌بندی تصاویر در صحنه‌های پیچیده معرفی می‌شوند.

۲ - ۱۰ دسته‌بندی تصاویر نامتعارف

اشتراک منابع اطلاعاتی مانند تصویر و توزیع آن در سطح جهانی با رشد سریع اینترنت و رسانه‌های اجتماعی تحت وب روز به روز در حال افزایش است. اگرچه این پدیده مزایای بسیاری مانند توسعه اشتراک محتوا در زمینه‌های آموزشی، خبری و غیره را داشته است، ولی نگرانی‌هایی را در این حوزه نیز ایجاد کرده است. از جمله این نگرانی‌ها می‌توان به انتشار تصاویر ناخواسته، توهین‌آمیز و نامتعارف[۱] در میان وب سایت‌های اینترنتی اشاره نمود که اهمیت فیلتر نمودن این تصاویر را نشان می‌دهد. همچنین

[۱] Pornography

بسیاری از سایت‌های اجتماعی درصدی اندک از تصاویر نامتعارف دارند. در صورت عدم وجود فیلترینگ هوشمند، سازمان‌های مرتبط مجبور به فیلتر نمودن کل سایت هستند که با توجه به استفاده وسیع کاربران از اینگونه سایت‌ها، احساس عدم رضایت در میان کاربران اینترنتی ایجاد می‌شود. در واقع به کمک فیلترینگ هوشمند می‌توان دسترسی کاربران به بسیاری از سایت‌ها را میسر نمود و تنها دسترسی به اطلاعات نامتعارف را محدود کرد.

بر اساس آمار ارائه شده، در سال ۲۰۱۱ میلادی دوازده درصد از سایت‌های موجود در شبکه اینترنت حاوی تصاویر نامتعارف هستند [۵۴]. بر اساس این آمار، تصاویر نامتعارف منتشر شده در اینترنت توسط ۴۲/۷ درصد از کاربران اینترنتی مشاهده شده است و از این تعداد ۳۴ درصد از کاربران علاقه‌ای به دیدن این تصاویر نداشته‌اند. با وجود اینکه روش دقیقی برای محاسبه تعداد وب‌سایت‌های مرتبط به پورن و همچنین بازدیدکنندگان آن وجود ندارد اما می‌توان تعداد این وب‌سایت‌ها را بسیار بالا تخمین زد. جدول ۲-۳ نتایج تکان‌دهنده‌ای از گزارش ارائه شده در مرجع [۴۵] در این زمینه را نشان می‌دهد. گزارش‌های مطرح شده در مراجع [۵۵, ۵۶] به بررسی اثرات پورنوگرافی در جوامع پرداخته است و نتایج این گزارش نشان می‌دهد که ارتباط مستقیمی بین پورنوگرافی و جنایت یا رفتارهای پرخاشگرانه‌ی جنسی در جامعه وجود دارد. استفاده از اینترنت به عنوان یک مرجع مهم برای دانش‌آموزان و استفاده متناوب جوانان و نوجوانان از این منبع اطلاعاتی، همواره خطراتی را برای جامعه دارد و این افراد را در معرض خطر تصاویر نامتعارف قرار می‌دهد. با توجه به موارد ذکر شده، معمولاً محدودیت‌های قانونی برای دسترسی به تصاویر نامتعارف برای افراد زیر سن قانونی اعمال می‌شود .

جدول ۲-۳ آمارهای مرتبط به هزینه و تعداد افراد در دسترس به اطلاعات پورنوگرافی[۱][۵۷]

در هر ثانیه	
هزینه انجام شده برای پورن	$۳۰۷۵,۶۴
تعداد مشاهده‌کنندگان پورن	۲۸۲۵۸
تعداد افراد جستجو کننده پورن	۳۲۷

[۱] **پورنوگرافی** یا هرزه نگاری، تجسمی بی پرده از مسایل جنسی با هدف تحریک یا ارضاست که در قالب‌های مختلف از جمله کتاب، فیلم، اینترنت و مجله ارائه می‌شود. در اینجا تمرکز بر روی تصاویر می‌باشد

با توجه به گسترده بودن این اطلاعات اعمال محدودیت نمی‌تواند همه‌ی مشکلات در این زمینه را رفع نماید. در همین راستا، در سال‌های اخیر پژوهش‌های بسیاری در زمینه‌های پردازش تصویر برای طراحی سیستمی به منظور فیلترینگ هوشمند انجام شد. با توجه به ماهیت تصاویر و پیچیدگی‌های موجود در ساختار تصویر، تشخیص و جداسازی تصاویر نامتعارف به‌سادگی امکان‌پذیر نیست.

در بسیاری از کارهای انجام‌شده، از رنگ پوست و ویژگی‌های مرتبط با آن برای تشخیص تصاویر نامتعارف استفاده می‌شود. تشخیص پیکسل‌های مرتبط با رنگ پوست به عنوان اولین گام در شناسایی این نوع از تصاویر است اما به اندازه کافی برای شناسایی تصاویر نامتعارف مورد اطمینان نیستند. در سیستمی که برای تشخیص این نوع از تصاویر در گوگل راه‌اندازی شد، سرعت پردازش و دقت عمل، به خصوص برای حجم بالای داده‌ها مورد توجه است. در سیستم مذکور طبقه‌بند قدرتمندی بر پایه ۲۷ ویژگی استخراج شده از پوست و طبقه‌بند ماشین‌بردار پشتیبان استفاده شد. ولی با توجه به حجم بالای این نوع از داده‌ها، همچنان این تحقیق نیازمند پژوهش‌های بیشتر است [۵۸].

۲ - ۱۰ - ۱ کارهای انجام شده در تصاویر نامتعارف

در سال‌های اخیر تحقیق‌های بسیاری بر روی تشخیص تصاویر نامتعارف انجام شده است. در این تحقیق‌ها، روش‌های فیلترینگ تصاویر نامتعارف برپایه محتوا و فیلترینگ هوشمند جایگاه ویژه‌ای دارند. در این نوع از فیلترینگ، معمولاً از طبقه‌بندهای آماری استفاده می‌شود. در این طبقه‌بندی، تصاویر به دو دسته‌ی متعارف و نامتعارف تقسیم می‌شوند. به صورت کلی تحقیقات انجام شده برای شناسایی تصاویر نامتعارف به چهار دسته بر اساس ساختار بدن انسان، بازیابی تصویر، کلمات بصری، و ویژگی‌های مبتنی بر پوست تقسیم می‌شوند. در ادامه کارهای انجام شده با توجه به این تقسیم‌بندی معرفی و تشریح می‌شود.

روش‌های مبتنی بر ساختار بدن انسان از سال ۱۹۹۶ با تحقیق انجام شده در مرجع [۵۹] مطرح شد. در این نوع از روش‌ها، نواحی مرتبط با رنگ پوست تشخیص و در گام بعدی با توجه به اطلاعات ساختار

بدن و نواحی استخراجی از رنگ پوست، قسمت‌های مرتبط با بدن انسان شناسایی و در نهایت با توجه به نواحی استخراج شده از بدن، تصاویر نامتعارف تشخیص داده می‌شوند. با توجه به غیرصلب[۱]بودن بدن انسان و وجود حالت‌های مختلف از بدن در تصویر، به دست آوردن یک حالت جامع از مدل بدن انسان کار دشواری است. تاکنون مدل خوبی برای ساختار بدن انسان ارائه نشده است. از این‌رو، روش‌های مبتنی بر ساختار بدن انسان دقت کمی دارند [۶۰].

هدف از روش‌های مبتنی‌بر بازیابی[۲]تصویر، به دست آوردن بهترین انطباق تصویری برای شناسایی زیرکلاس‌های مجموعه داده است و از آنالیز مستقیم یک تصویر استفاده نمی‌شود. در این روش‌ها، تصاویر نامتعارف منطبق با تصویر ورودی از روی پایگاه داده استخراج می‌شوند. در صورتی که تعداد تصاویر نامتعارف انطباقی بیشتر از حد آستانه تعیینی باشد، تصویر ورودی به عنوان تصویر نامتعارف معرفی می‌شود. شناسایی در این روش‌ها، تا حد زیادی وابسته به پایگاه داده است. با توجه به فراوانی تصاویر متعارف و نامتعارف در شکل‌های مختلف، ساخت پایگاه داده کامل بسیار دشوار است. برای بهبود دقت شناسایی در اینگونه روش‌ها، نیاز به تعداد بالای نمونه‌ها است که زمان شناسایی را در مرحله آزمون کاهش می‌دهد. در سال‌های اخیر شناسایی تصاویر مبتنی بر کلمات بصری بسیار مورد توجه محققان قرار گرفته که این روش محبوبیت زیادی در طبقه‌بندی اشیاء به دست آورده است. با توجه به مطالب ارائه شده در بخش مجموعه کلمات، کلمات بصری به قطعات کوچکی از تصویر اشاره می‌کنند که نوعی از اطلاعات مرتبط با ویژگی‌هایی مانند رنگ، شیء یا بافت را دارا هستند. محققان روشی برای آنالیز تصویر با الهام‌گیری از آنالیز محتوای متن ارائه کرده‌اند. در این روش، یک تصویر به عنوان ترکیبی از کلمات در نظر گرفته می‌شود و روش آنالیز متنی مورد استفاده در تفسیر معنایی[۳]برای شناسایی تصاویر نامتعارف استفاده می‌شود. در این روش‌ها کلمات بصری برای توصیف محتوای معنایی تصویر استخراج می‌شوند. سپس مدل مجموعه کلمات برای شناسایی تصاویر نامتعارف معرفی می‌شود [۶۱].کلمات

---

[۱] Non-rigid

[۲] Retrieval

[۳] Semantic Annotation

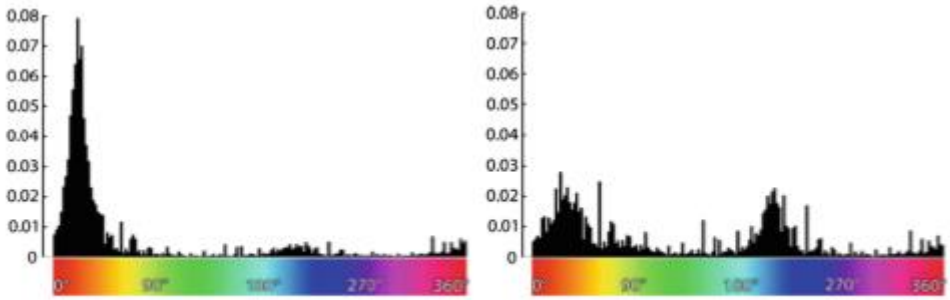
بصری و اطلاعات معنایی از برچسب‌های مرتبط[۱] برای تخمین امتیاز ارتباط بین برچسب‌های موجود در تصاویر استفاده می‌شود. مدل مجموعه کلمات برای شناسایی تصاویر نامتعارف در مرجع [۶۲] استفاده شد. در این تحقیق توصیفگر SIFT برای تولید واژگان بصری استفاده شده است و برای شناسایی تصاویر نامتعارف از طبقه‌بند ماشین بردار پشتیبان استفاده شد. در مرجع [۶۳] از اطلاعات متن و مجموعه کلمات بصری برای شناسایی تصاویر نامتعارف استفاده شد. در این تحقیق بجای توصیفگر SIFT از روش SURF[۲] استفاده شد.

در روش‌های مبتنی بر ویژگی‌های پوستی، شناسایی تصاویر مانند یک مسئله طبقه‌بندی است. در این روش‌ها، ویژگی‌های رنگی و بافتی از پوست استخراج می‌شوند و در گام بعدی از طبقه‌بندهای الگو برای شناسایی و طبقه‌بندی تصاویر استفاده می‌شوند. بسیاری از کارهای انجام شده در سال‌های اخیر مرتبط با این روش بوده است. کارهای انجام شده در این زمینه، به سه دسته کلی مبتنی بر اطلاعات رنگ، اطلاعات شکل و توصیفگرهای ویژگی‌های محلی تقسیم می‌شوند. روش‌های مبتنی‌بر ویژگی‌های رنگی پیکسل، معمولاً برای شناسایی نواحی پوست استفاده می‌شوند. در مرجع [۶۴] بیش از دویست هزار تصویر از سایت‌های متعارف و نامتعارف انتخاب شد و نشان داد که مقدار فام[۳]در این دو دسته از تصاویر بسیار متفاوت هستند (شکل ۲-۳۰). هر پیکسل از تصویر با توجه به مدل رنگ می‌تواند برای شناسایی پیکسل‌های پوستی و غیر پوستی استفاده شود. این روش‌ها معمولاً توسط طبقه‌بندهای ماشین بردار پشتیبان، شبکه‌های عصبی پرسپترون و درخت تصمیم دسته‌بندی گردند.

دسته‌ی دوم روش‌های فیلترینگ هوشمند مبتنی بر پوست، بر پایه اطلاعات شکل هستند. این روش‌ها بر تشخیص نواحی پوستی تکیه دارند. همچنین در این روش‌ها از تعدادی ویژگی‌های مبتنی‌بر توصیفگر نواحی تصویر استفاده شد. ویژگی‌های مرتبط به شکل را می‌توان در پنج دسته، ویژگی‌های مبتنی‌بر لبه، گشتاورها[۴] محدودیت‌های هندسی[۵] و قطعه‌بندی رنگ[۶] دسته‌بندی نمود [۶۵].

---

[۱] Associated Tags
[۲] Speeded up Robust Features
[۳] Hue
[۴] Moments
[۵] Geometric Constraints
[۶] Color Segments



شکل ۲-۳۰ مقایسه بین نتایج به دست آمده از توزیع مقدار فام (Hue) در تصاویر استخراج شده از وب‌سایت‌های متعارف و نامتعارف [۶۴]، مقدار نرمال شده فام با زاویه بین ۰ تا ۳۶۰ درجه با توجه به تعریف فضای رنگی HSV محاسبه شده است.الف) تصاویر متعارف، ب) تصاویر نامتعارف

دسته‌ی سوم روش‌های فیلترینگ هوشمند مبتنی بر پوست، از بردار ویژگی برای توصیف یک ناحیه از تصویر استفاده می‌کنند. در این روش‌ها نقاط کلیدی با نمونه‌گیری تنک از تصویر انتخاب می‌شوند و از این نقاط به عنوان ویژگی برای شناسایی تصاویر مشابه استفاده شد. در کارهای انجام شده توسط این روش می‌توان به SIFT و PLSA[۱] اشاره نمود [۶۶]. چالش اصلی در این نوع از روش‌ها، استخراج ویژگی مؤثر و مفید برای جداسازی تصاویر متعارف از نامتعارف است. در مرحله‌ی اول از تصاویر ویژگی استخراج می‌شود و در گام بعدی طبقه‌بندی انجام می‌شود که عملکرد طبقه‌بند به مرحله استخراج ویژگی بسیار وابسته است.

در تصاویری که شامل تصاویر نامتعارف می‌باشند معمولا تعداد اشیاء موجود و تعداد کلاس و پیچیدگی صحنه پایین می‌باشد. ولی در بسیاری از کاربردها معمولا پیچیدگی صحنه بالا است و تعداد کلاس‌های صحنه افزایش می‌یابد. در همین راستا در ادامه این فصل، روش‌های موجود در دسته‌بندی صحنه‌های پیچیده‌تر از لحاظ تعداد صحنه‌ها و تعداد اشیاء معرفی می‌شود.

[۱] probabilistic Latent Semantic Analysis (pLSA)

۲ - ۱۱ دسته‌بندی صحنه‌های پیچیده

در سال‌های اخیر، دسته‌بندی صحنه‌ی تصویر به عنوان یکی از موضوعات مهم در بینایی کامپیوتر و یادگیری ماشین مطرح شده است. به دست آوردن خودکار اطلاعات معنایی تصویر یک امر ضروری بوده و کاربردهای فراوانی در دنیای واقعی دارد. روش‌های زیادی برای این منظور در سال‌های اخیر معرفی شده است. در روش‌های قدیمی‌تر، هر تصویر به‌عنوان یک شی مجزا در نظر گرفته شده و با ویژگی‌های سطح پایین عمل دسته‌بندی انجام می‌شود. این روش‌ها به طور معمول، فقط برای دسته‌بندی تعداد کمی از دسته‌های صحنه استفاده شده و نمی‌تواند برای دسته‌بندی تصاویر دنیای واقعی استفاده شود. در سال‌های اخیر برای این دسته‌بندی از نمایش سطح بالای تصویر با چندین نواحی از تصویر استفاده شد که عملکرد خوبی را در این کاربرد از خود نشان داده است. استفاده از مدل‌های موضوعی ساخته شده توسط متغیرهای پنهان در روش‌های نوین برای دسته‌بندی صحنه مطرح می‌باشند که در سال‌های اخیر مورد توجه بسیار از محققان قرار گرفته است. در این روش‌ها دسته‌بندی تصویر برپایه معناشناسی تصویر انجام می‌شود. این روش‌ها برای مواردی که تعداد کلاس‌های صحنه بالا است مورد استفاده قرار می‌گیرد. به همین منظور در این فصل به معرفی روش‌های دسته‌بندی تصویر برپایه مدل‌های موضوعی پرداخته شده است.

یکی از روش‌های غالب در دسته‌بندی صحنه/تصویر بر پایه ویژگی‌های محلی می‌باشد. در ابتدا نقاط مهم و جذاب از تصاویر برپایه آشکارسازها استخراج می‌شود و سپس توصیفگرهای محلی مانند SIFT برای توصیف این نقاط استخراج می‌شوند. نکته مهم در این نوع از ویژگی‌ها اندازه‌گیری شباهت بین دو تصویر در مجموعه‌هایی از توصیفگرهای محلی می‌باشد که با این وجود به دلیل قدرت بالای این ویژگی‌های محلی این اصل نادیده گرفته می‌شود. در دهه‌های گذشته، رویکردهای مبتنی بر مدل مجموعه کلمات و الگوریتم کدگذاری تنگ نتایج قابل توجهی در بسیاری از کارهای چالش برانگیز به دست آورده‌اند. این دو نوع رویکرد، ویژگی‌های سطح متوسط را نمایش می‌دهند [۶۷]. عیب اصلی مدل مجموعه کلمات و کدگذاری تنک ، خطاهای کوانتیزه کردن و از دست دادن اطلاعات ساختاری از

ویژگی‌های محلی است. تلاش‌های زیادی برای رفع اشتباهات کوانتیزه شده ناشی از خسارت اطلاعات مکانی توصیفگرهای محلی صورت گرفته است. با توسعه روش‌های شناسایی شی روش‌های امیدوار کننده برای تحلیل تصاویر برپایه ویژگی‌های سطح بالا به وجود آمده است. یک الگوریتم شناخته شده روش نمایش مخزن اشیاء [۶۸] می‌باشد. در این مرجع نشان داده شده که ویژگی‌های سطح پایین نمی‌توانند ابهامات موجود در صحنه تصویر را کنترل کنند. در همین راستا باید از روش‌های استفاده نمود که بر پایه ویژگی‌های معنایی سطح بالا باشند که در ادامه این روش‌ها معرفی می‌شود.

۲ - ۱۲ کارهای انجام شده در دسته‌بندی صحنه‌های پیچیده بر پایه مدل‌سازی موضوعی

نمایش مجموعه کلمات تا حدود زیادی کاربرد مدل‌های موضوعی نهفته را تسهیل داده است و از ادبیات درک اسناد برای شناسایی صحنه استفاده کرده است. برای نمونه می‌توان به روش تجزیه و تحلیل معنایی نهفته آماری [۶۹]، تخصیص پنهان دیریکله[۱](LDA) [۷۰] اشاره نمود. برای نمونه در مرجع [۷۱] از روش LDA برای شناسایی دسته‌های صحنه استفاده شد. در مرجع [۷۲] یک مدل موضوعی سلسله‌مراتبی[۲]برای اشیاء بر پایه بخش‌ها[۳]و شناسایی دسته صحنه معرفی شد. در روش پیشنهادی، بخش‌ها می‌توانند در میان دسته‌های بصری مختلف به اشتراک گذاشته شوند. مدل پیشنهادی، یک تصویر از صحنه را در سه لایه سلسله‌مراتبی نمایش می‌دهد. بالاترین لایه مربوط به یک صحنه (برای نمونه "خیابان"، "ساحل"، "جنگل" و ..)، لایه میانی[۴]مرتبط به مجموعه‌ای از عناصر[۵]صحنه (مانند "ساختمان"، "آسمان"، "کوه" و غیره) و لایه پایین[۶] مرتبط به مجموعه از ویژگی‌های استخراج شده

---

[۱] Latent Dirichlet Allocation (LDA)
[۲] Hierarchical
[۳] Part Based
[۴] Themiddle
[۵] Elements
[۶] Bottom Level

از وصله‌های[۱]محلی تصویر است. بنابراین این روش یک صحنه‌ی تصویر را به وسیله عناصر صحنه و کلمات بصری بدون نیاز به برچسب‌گذار عناصر صحنه مدل می‌کند.

اگرچه این روش‌ها در دسته‌بندی بدون‌نظارت، مؤثر است ولی از اطلاعات بانظارت استفاده مؤثر نشده است. در سال‌های اخیر محققان بسیاری به مدل‌سازی موضوعی تصویر بانظارت برای شناسایی بصری پرداخته‌اند [۷۳, ۷۴]. با توجه به طراح‌های[۲]مختلف که از اطلاعات بانظارت استفاده شده، مدل‌سازی موضوعی بانظارات را می‌توان در دوسته، مدل‌سازی پایین‌دست[۳]و بالادست[۴]تقسیم نمود [۷۵]. برای مدل‌سازی پایین‌دست، پاسخ متغیرهای بانظارت از متغیرها انتساب موضوع (برای نمونه sLDA [۷۵]) تولید شده‌اند. برای مدل‌سازی بالادست، متغیرهای پاسخ به صورت مستقیم یا غیرمستقیم از طریق متغیرهای موضوعی نهفته (برای نمونه SidcLDA [۷۴]) تولید شده‌اند که این روش‌ها برای تحقیقات بینایی انسان[۵]مورد توجه محققین قرار گرفته است. هر دو روش مدل‌سازی موضوعی از بخش‌بندی برای شناسایی صحنه استفاده می‌کنند. برای نمونه مرجع [۳]، روش sLDA را برای دسته‌بندی و برچسب‌زنی تصویر به صورت هم‌زمان توسعه داده است. همچنین در مرجع [۷۳] روش SidcLDA را به وسیله مدل کردن یک آرایش فضایی[۶]از قسمت‌های اشیاء یا نواحی صحنه برای شناسایی بصری توسعه داده است. به عبارت دیگر، بسیاری از مدل‌های موضوعی قدیمی از نمایش مجموعه کلمات به‌صورت مستقیم استفاده می‌کردند که این روش‌ها محتوای فضایی[۷]ویژگی‌های محلی را نادیده می‌گرفتند. در تحقیقات قبلی [۳, ۷۳, ۷۶, ۷۷] نشان داده شده است که اطلاعات محتوا[۸]مانند تعامل[۹]روابط[۱۰]بین ویژگی‌های محلی، نواحی تصویر و شیء/صحنه، اطلاعات مفیدی[۱۱]در رفع ابهام[۱۲]کلمات بصری فراهم می‌کند که اغلب این روش‌ها منجر به بهبود عملکرد سیستم شناسایی می‌شود.

---

[۱] Patches
[۲] Schemes
[۳] Downstream
[۴] Upstream
[۵] Human Vision
[۶] Spatial Arrangement
[۷] Spatial Context
[۸] Contextual
[۹] Interaction
[۱۰] Relationships
[۱۱] Beneficial
[۱۲] Disambiguating

دو روش برای مدل‌سازی اطلاعات محتوا وجود دارد، یکی از این روش‌ها برای مدل‌سازی لایه‌های فضایی، از تکه‌های تصویر یا اشیا، استفاده کرده‌اند. روش دیگر به مدل‌سازی رابطه‌ی بین تکه‌ها و اشیای تصویر همسایه تمرکز دارد. این روش‌ها به عنوان محتوای محلی[۱] که توسط فاصله بین تکه‌های تصویر یا اشیای آن محاسبه می‌شود نام‌گذاری می‌شود. برای مدل محتوای سراسری، اطلاعات موقعیت از تکه‌های تصویر (برای نمونه ویژگی محلی) استفاده شد [۸, ۷۳, ۷۸]. شکل اشیاء یا موقعیت متقابل از قطعات در [۷۸] نمایش داده شد. انطباق هرم فضایی[۲][۸] و نوع افتراقی[۳] آن [۷۹] عملکرد بسیار خوبی هم در شناسایی شیء و صحنه از خود نشان داده‌اند. مرجع [۸۰] یک سری از مرتبه‌های[۴] بسته ویژگی‌ها را توسط نگاشت ویژگی‌های محلی به خط‌ها و منحنی‌های[۵] متفاوت تولید کرده است. در مرجع [۷۶] یک محتوای معنایی[۶] چندگانه مخصوص هیستوگرام‌های مجموعه کلمات را برای نمایش یک تصویر ساخته و به نمایندگی یک تصویر در ساختار سلسله‌مراتبی نمایش داده شد. در تعدادی از کارهای گذشته [۷۳, ۷۷] نشان داده شد که مدل از محتوا بر اساس "صحنه/شیء"،"عناصر صحنه/قسمت‌های اشیاء" و"تکه‌های تصویر" بسیار در شناسایی با تجزیه[۷] صحنه قوی‌تر[۸] است.

مدل محتوای محلی، رابطه‌ی تکه‌های همسایه‌ی تصویر را بررسی می‌کند. مرجع [۸۱] روش sLDA را برای کشف کلاس‌های اشیاء از مجموعه‌ای از تصاویر پیشنهاد داده است. در sLDA، کلمات بصری (به عنوان مثال تکه‌های چشم و تکه‌های بینی) که اغلب در تصاویر مشابه رخ می‌دهند و فضای نزدیک به هم دارند را به عنوان یک موضوع (به‌عنوان نمونه صورت) خوشه‌بندی کرده است. مشابه هیستوگرام محتوای اشیاء، یک هیستوگرام خاص با فاصله تخمینی در مقیاس لگاریتمی برای ویژگی‌های خاص معرفی شده در [۸۲] ارائه شد. در مرجع [۸۳] یک مدل موضوعی با در نظر گرفتن این نکته که کلمات

---

[۱] Local Context
[۲] Spatial Pyramid Matching (SPM)
[۳] Discriminative
[۴] Ordered
[۵] Circles
[۶] Semantic
[۷] Parsing
[۸] Robust

بصری در ناحیه محلی منسجم در یک موضوع مشابه اشتراک دارند معرفی کرد. این روش توانسته یک مدل محتوا محلی بسیار مؤثر را به وجود بیاورد.

همان‌طور که قبلاً ذکر شد مدل‌سازی موضوعی بر پایه LDA یک چارچوبی از انتخاب‌ها برای مقابله با داده‌های چندین‌حالته[۱] مانند برچسب‌زنی تصویر می‌باشد. یکی دیگر از روش‌های مشهور که در سال‌های اخیر برای مدل‌سازی داده‌های چندحالته مورد استفاده قرار گرفته، شبکه‌عصبی عمیق[۲] مانند ماشین بلتزمن عمیق[۳](DBM) است. اخیراً، نوع جدیدی از مدل موضوعی با عنوان تخمین‌گر توزیع‌شده خودرگرسیو عصبی اسناد[۴](DocNADE) پیشنهاد شد و در [۸۴] ادعا شد که بهترین عملکرد را در کارهای نوین[۵] برای مدل‌سازی متن اسناد دارد. در مرجع [۱] نشان داده شد که چطور می‌توان این روش و توسعه‌یافته این مدل را بر روی داده چندین حالته مانند استفاده هم‌زمان از دسته‌بندی تصویر و برچسب‌زنی استفاده کرد. در گام نخست، یک سیستم توسعه‌یافته بانظارت از DocNADE با عنوان SupDocNADE پیشنهاد شد. این گام قدرت جداسازی را توسط آموزش دادن ویژگی‌های موضوعی پنهان افزایش داد. همچنین نشان داد چطور از این روش برای اشتراک نمایش‌های مختلف از اطلاعات کلمات بصری تصویر، کلمات برچسب‌زنی و برچسب کلاس استفاده شود. روش پیشنهادی بر روی مجموعه داده LabelME و UIUC-Sport مورد آزمایش قرار گرفت و نشان داده شد که این روش قابل مقایسه با روش‌های دیگر مدل‌سازی موضوعی می‌باشد. در گام دوم یک مدل عمیق توسعه یافته از مدل خود ارائه داد و یک راه مؤثر برای آموزش مدل عمیق ارائه کرد. به دلیل دقت بالای این روش، در ادامه این روش به تفصیل بیان می‌شود.

---

[۱] Multimodal

[۲] Deep Neural Networks

[۳] Deep Boltzmann Machine (DBM)

[۴] Document Neural Autoregressive Distribution Estimator (DocNADE)

[۵] state-of-the-art

۲ - ۱۲ - ۱ تخمین‌گر توزیع‌شده خودرگرسیو عصبی اسناد

یکی از تحقیقاتی که در سال‌های اخیر نتایج بسیار خوبی را در مدل‌سازی موضوعی تصویر از خود نشان داد و بسیار مورد توجه قرار گرفت مرجع [۱] است. در اینجا به تفصیل این روش معرفی می‌شود. در این مرجع از مدل DocNADE ارائه شده توسط [۸۴] استفاده شد. DocNADE برای مدل‌سازی اسناد از کلمات واقعی استفاده می‌کند که این کلمات به برخی از واژگان از پیش تعریف شده تعلق دارند. برای مدل کردن داده‌ها بر پایه تصویر، فرض شد که در ابتدا تصاویر به مجموعه کلمات تبدیل شده‌اند. یک روش استاندارد برای آموزش واژگان کلمات بصری، استفاده از خوشه‌بندی k-means بر روی توصیفگرهای SIFT مجموعه تصاویر می‌باشد. در این مقاله پیرو مقاله [۳] از ویژگی‌های SIFT استخراج شده در ۱۲۸ بعد برای استخراج کلمات بصری استفاده شد. اندازه گام و وصله از SIFT متراکم استخراج شده در مجموعه ۸ و ۱۶ می‌باشد. ویژگی‌های SIFT متراکم از مجموعه آموزش در ۲۴۰ خوشه رقمی‌سازی شد و برای ساختار کلمات بصری واژگان از k-mean استفاده شد. هر تصویر در شبکه مشبک ۲*۲ برای استخراج اطلاعات موقعیت فضایی تقسیم شد. در اینجا با توجه به کارهای انجام شده، ۹۶۰=۲۴۰*۲*۲ جفت ناحیه/کلمه بصری متفاوت تولید می‌شود. برای استفاده از اطلاعات فضایی اشاره شد که این خصیصه نقش مهمی را در درک تصویر ایفا می‌کند. در واقع از موقعیت کلمات بصری که دستاورد قابل توجهی در عملکرد داشته استفاده شده است.

هر تصویر توسط مجموعه کلمات بصری $v = [v_1, \dots, v_D]$ توصیف می‌شود که هر v_i ایندکس نزدکترین خوشه k-means در iامین توصیفگر استخراج شده از تصویر می‌باشد و D تعداد توصیفگرهای استخراج شده است. مدل DocNADE احتمال مشترک کلمات بصری $p(v)$ را به صورت رابطه (۲-۲۰) تعریف می‌کند.

$$p(v) = \prod p(v_i|v_{<i}) \qquad (۲۰-۲)$$

بجای هر شرط، $p(v_i|v_{<i})$ که $v_{<i}$ زیربرداری که شامل همه‌ی v_j که $j < i$ می‌باشد، مدل‌سازی می‌شود. اشاره شد که معادله (۲-۲۰) برپایه احتمال زنجیره قوانین برای هر توزیع درست می‌باشد. فرض

اصلی توسط DocNADE در شکل شرطی ساخته شد و به‌طور خاص در DocNADE فرض می‌شود که در هر شرایطی توسط شبکه عصبی پیشرو، مدل و یاد گرفته شود.

یک احتمال می‌تواند برای مدل کردن $p(v_i|v_{<i})$ توسط معماری زیر تعریف شود (رابطه (۲-۲۱)):

$$h_i(v_{<i}) = g\left(c + \sum_{k<i} W_{:,v_k}\right), p(v_i = w|v_{<i}) = \frac{\exp(b_w + V_{w,:}h_i(v_{<i}))}{\sum_{w'} \exp(b_{w'} + V_{w',:}h_i(v_{<i}))} \quad (۲۱-۲)$$

در اینجا $g(.)$ یک تابع فعال‌ساز غیرخطی مبتنی بر عناصر[۱] می‌باشد؛ $W \in R^{H\times K}$ و $V \in R^{K\times H}$ ماتریس پارامتر مشترک، $c \in R^N$ و $b \in R^K$ بردار پارامتر بایاس و H، K تعداد واحدهای پنهان (موضوع) و اندازه واژگان می‌باشد.

محاسبه توزیع $p(v_i = w|v_{<i})$ در معادله (۲-۲۱) به زمان خطی K نیازمند می‌باشد. در عمل از آنجایی‌که این روش باید برای هر D کلمات بصری v_i محاسبه شود بسیار زمان‌بر است. برای غلبه بر این مشکل مقاله [۸۴] استفاده از روش درخت باینری متوازن را برای تجزیه محاسبه شرطی و به دست آوردن زمان لگاریتم برحسب K معرفی کرد. برای این منظور به صورت تصادفی همه‌ی کلمات بصری به برگ‌های متفاوت در درخت باینری انتساب داده شده است. با توجه به این درخت، احتمال یک کلمه با احتمال به دست آوردن برگ مرتبط از ریشه مدل می‌شود. هر احتمال انتقال چپ/راست در درخت باینری توسط یک مجموعه از رگرسیون لجستیک باینری[۲] محاسبه شده از لایه مخفی $h_i(v_{<i})$، به عنوان ورودی، مدل می‌شود. احتمال یک کلمه داده شده به وسیله ضرب احتمالات از هر انتخاب چپ/راست از مسیر درخت مرتبط به دست می‌آید.

به صورت خاص، $l(v_i)$ توالی از نودهای درخت در مسیر از ریشه تا برگ v_i و $\pi(v_i)$ توالی از انتخاب‌های چپ/راست باینری در نودهای داخلی متعلق به مسیر آن می‌باشد. برای مثال $l(v_i)_1$، همیشه به عنوان نود ریشه از درخت باینری و $\pi(v_i)_l$ درصورتی‌که برگ کلمه v_i در زیر درخت چپ باشد صفر خواهد

[۱] Element-Wise

[۲] Binary Logistic Regressors

بود در غیر این صورت یک خواهد بود. $V \in R^{T\times H}$ حالا ماتریسی شامل وزن‌های رگرسیون لجستیک و $b \in R^T$ برداری شامل بایاس می‌باشد که T در اینجا تعداد نودهای داخلی در درخت باینری و H تعداد واحدهای پنهان می‌باشد. احتمال $p(v_i = w|v_{<i})$ حالا به‌صورت رابطه (۲-۲۲) مدل می‌شود:

$$p(v_i = w|v_{<i}) = \prod_{k=1}^{|\pi(v_i)|} p(\pi(v_i)_k|v_{<i}) \qquad (۲۲-۲)$$

$$p(\pi(v_i)_k = 1|v_{<i}) = sigm\left(b_{l(v_i)_m} + V_{l(v_i)_m,:}.\, h_i(v_{<i})\right) \qquad (۲۳-۲)$$

که در اینجا نود داخلی، خروجی‌های رگرسیون لجستیک و تابع سیگموئید (رابطه (۲-۲۳)) هستند. با استفاده از درخت متوازن، زمان محاسبه رابطه (۲-۲۲) را در خروجی رگرسیون لجستیک با $O(\log K)$ تضمین می‌کند.

با محاسبه رابطه (۲-۲۱)، (۲-۲۲) و (۲-۲۳) می‌توان احتمال $p(v) = \prod_{i=1} p(v_i|v_{<i})$ را برای هر سند در DocNADE محاسبه کرد. آموزش پارامترهای $\theta = \{W, V, b, c\}$ از DocNADE با استفاده از منفی درست‌نمایی لگاریتمی[۱]مجموعه آموزش به همراه شیب نزولی اتفاقی[۲]بهینه می‌شود. هنگامی که یک مدل آموزش داده شد، یک نمایش پنهان می‌تواند از یک سند جدید v^* توسط رابطه (۲-۲۴) به دست بیاید.

$$h_y(v^*) = g\left(c + \sum_i^D W_{:,v_i^*}\right) \qquad (۲۴-۲)$$

این نمایش می‌تواند برای یک دسته‌بند استاندارد به منظور بهبود هر وظیفه‌ی بینایی ماشینی بانظارت تغذیه شود. رابطه (۲-۲۲) و (۲-۲۳) نشان می‌دهد که احتمال شرطی هر کلمه v_i نیازمند محاسبه موقعیت وابسته لایه پنهان $h_i(V_{<i})$ می‌باشد، که یک نمایش خروجی از مجموعه کلمات بصری قبل $(v_{<i})$ استخراج شده، می‌باشد. از آنجایی که محاسبه $h_i(v_{<i})$ به صورت متوسط در زمان O(HD)

[۱] Log-Likelihood

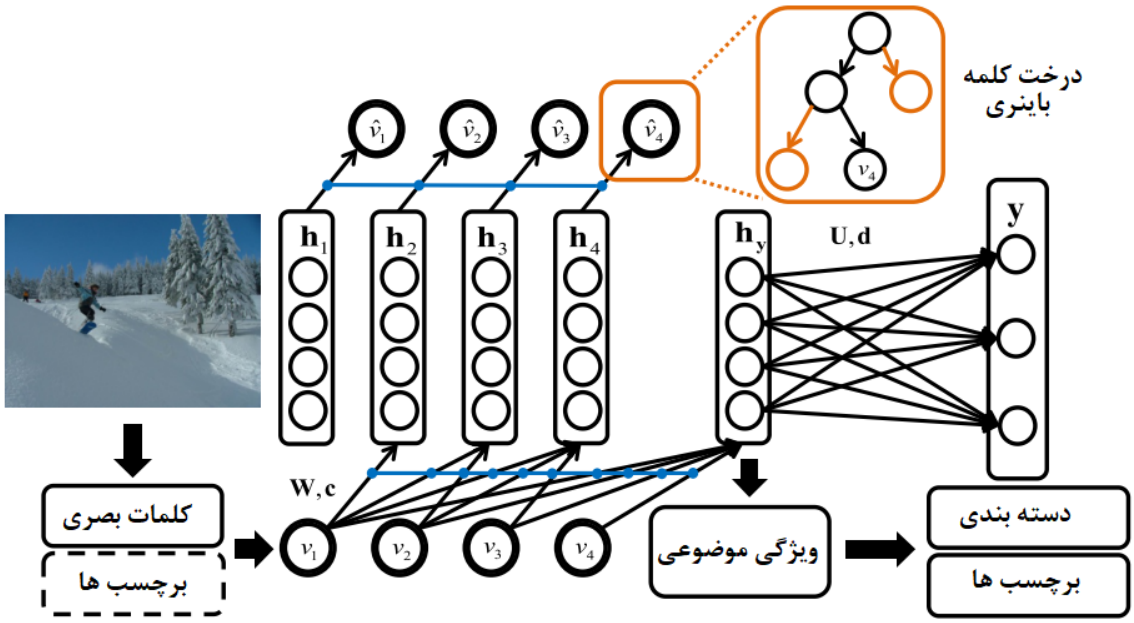
[۲] Stochastic

می‌باشد و D لایه پنهان $h_i(v_{<i})$ برای محاسبه وجود دارد پس یک روش بیز برای محاسبه همه لایه‌های پنهان در حدود $O(HD^2)$ زمان نیاز دارد.

مقادیر لایه‌های بعدی توسط رابطه (۲-۲۵) محاسبه می‌شود:

$$h_{i+1}(v_{<i+1}) = g\left(c + \sum_{k<i+1} W_{:,v_k}\right) = g(W_{:,v_i} + c + \sum_{k<i} W_{:,v_k}) \qquad (۲۵-۲)$$

با به کارگیری این اصل که ماتریس وزن W در همه‌ی شرایط انتقال خطی $c + \sum_{k<i} W_{:,v_k}$ دارد، می‌توان برای محاسبه $h_{i+1}(v_{<i+1})$ از محاسبه لایه‌های مخفی قبلی $h_i(v_{<i})$ ، مجدد استفاده نمود. با این روش، محاسبه همه‌ی لایه‌های مخفی $h_i(v_{<i})$ پی‌درپی از i=1 تا i=D در زمان $O(HD)$ می‌باشد. در نهایت از آنجایی‌که محاسبه پیچیدگی از هر $O(\log K)$ رگرسیون لجستیک در رابطه برابر با $O(H)$ است، در مجموع پیچیدگی محاسباتی برای محاسبه $p(v_i = w|v_{<i})$ برابر با $O(\log K\, HD)$ می‌باشد. در ادامه به تشریح روش پیشنهادی برای دسته‌بندی و برچسب‌گذاری می‌پردازیم (شکل ۲-۳۱). در [۱] تخمین‌گر توزیع‌شده خودرگرسیو عصبی اسناد با الهام‌گیری از DocNADE به‌صورت هم‌زمان به برچسب‌گذاری و کلاس‌بندی تصویر پرداخته است که روش پیشنهادی یک توسعه بانظارت روش DocNADE با عنوان SupDocNADE می‌باشد. در این روش از اطلاعات برچسب کلاس در مجموعه آموزش برای یادگیری ویژگی‌های پنهان افتراقی به منظور دسته‌بندی استفاده می‌شود. همچنین در ادامه، طریقه استفاده از موقعیت مکانی کلمات بصری و چگونگی به کارگیری برچسب‌گذاری به همراه دسته‌بندی در روش پیشنهادی بیان می‌شود.



شکل ۲-۳۱ نمایش یک لایه مخفی مدل SupDocNADE برای داده‌های تصویر چندحالته [۱]. کلمات بصری، کلمات برچسب‌گذاری شده و برچسب کلاس y به عنوان $p(v,y)=\prod_i p(v_i|v_1,\dots,v_{i-1})$ مدل شده‌اند. تمام شرایط $p(y|v)$ و $p(v_i|v_1,\dots,v_{i-1})$ با استفاده شبکه‌عصبی با اشتراک وزن مدل شده‌اند.

در مرجع [۱] بیان شد که که یادگیری ویژگی تصویر با به کارگیری مدل‌های موضوعی بدون ناظر مانند LDA می‌تواند عملکرد بدتری از آموزش دادن یک دسته‌بندی به صورت مستقیم بر روی کلمات بصری (مانند استفاده کردن از یک کرنل مناسب مانند کرنل هرمی) داشته باشد. به طور خاص، با توجه به تصویر $v=[v_1,v_2,\dots,v_D]$ و برچسب کلاس آن‌ها $y\in\{1,2,\dots,C\}$، SupDocNADE توزیع مشترک کامل را به صورت رابطه (۲-۲۶) مدل می‌کند:

$$p(v)=p(y|v)\prod_{i=1}^{D}p(v_i|v_{<i}) \qquad (۲۶-۲)$$

در DocNADE هر شرط با استفاده از شبکه‌عصبی مدل شد. در [۱] برای $p(v_i|v_{<i})$ از همان ساختار استفاده شده است فقط باید روشی برای $p(y|v)$ معرفی شود. از آنجایی‌که از $h_y(v)$ برای نمایش تصویر برای دسته‌بندی استفاده شد. مدل $p(y|v)$ به عنوان خروجی رگرسیون لجستیک چندکلاسه از $h_y(v)$ به صورت رابطه (۲-۲۷) محاسبه می‌شود.

$$p(y|v)=Softmax\big(d+Uh_y(v)\big)_y \qquad (۲۷-۲)$$

در اینجا $Softmax(a)_i = \exp(a_i) / \sum_{j=1}^{C} \exp(a_j)$، $d \in R^C$ بردار پارامتر بایاس در لایه بانظارت و $U \in R^{C \times H}$ ماتریس اتصال بین لایه پنهان h_y و برچسب کلاس می‌باشد.

به عبارت دیگر، $p(y|v)$ به عنوان شبکه عصبی چندکلاسه منظم با گرفتن ورودی مجموعه کلمات v مدل‌سازی شد. تفاوت اساسی این روش با یک شبکه عصبی این می‌باشد که بعضی از پارامترهای آن (یعنی پارامترهای واحد پنهان W ,C) در مدل‌سازی کلمات بصری شرطی $p(v_i|v_{<i})$ استفاده شده اند.

ماکزیمم درست‌نمایی آموزش داده شده از این مدل، مشابه رابطه (۲-۲۸) توسط مینیم‌کردن منفی لگاریتم درست‌نمایی به دست می‌آید.

$$-\log p(v|y) = -\log p(y|v) + \sum_{i=1}^{D} -\log(v_i|V_i) \qquad (۲-۲۸)$$

از تمام تصاویر مجموعه آموزش متوسط گرفته می‌شود. این روش به عنوان یادگیر مولد شناخته شده است. ترم اول یک ترم افتراقی محض[1](کامل) می‌باشد. درصورتی‌که ترم دوم، بدون ناظر می‌باشد و می‌توان به عنوان تنظیم‌کننده استفاده شود که یک راه‌حل می‌باشد که همچنین ساختار آماری بدون ناظر کلمات بصری را توضیح می‌دهد تشویق می‌کند. در عمل این تنظیم‌کننده می‌تواند راه‌حل را به جوابی خیلی دور از راه‌حل افتراقی که به خوبی تولید شده است، بایاس کند. در [۱] پیشنهاد شد مشابه کارهای گذشته بر روی ترکیب یادگیری افتراقی/ مولدی از رابطه (۲-۲۹) استفاده شود که در آن به عملکرد مولد وزن اهمیت داده می‌شود.

$$-\log p(v,y) = -\log p(y|v) + \lambda \sum_{i=1}^{D} -\log(v_i|V_i) \qquad (۲-۲۹)$$

در اینجا λ به عنوان یک تنظیم‌کننده ابرپارامتر رفتار می‌کند. آموزش دادن بر روی میانگین مجموعه آموزش از رابطه (۲-۲۹) به وسیله گرادیان نزولی اتفاقی انجام شد. از پس‌انتشار برای محاسبه مشتق

[1] Purely

پارامترها استفاده شد. از آنجایی که در DocNADE منظم، برای محاسبات آموزش و گرادیان آن باید یک ترتیب از ورودی‌ها تعریف شود. همچنین از آنجایی که می‌توان یک مسیر دلخواه از ترتیب کلمات (برای نمونه، از راست به چپ، بالای به پایین در تصویر) تعریف کرد. در همین راستا مشابه کار انجام شده در مرجع [۸۴] به صورت تصادفی کلمات قبل از هر بروز رسانی توسط گرادیان تصادفی تغییر می‌کنند. مفهوم این کار این است که مدل برای هر ترتیب از کلمات در V به طور مؤثر به یک مدل استنتاجی در هر شرایط $p(v_i|v_{<i})$ آموزش ببیند. همچنین با این روش با بیش‌برازش مقابله شد و مدل پیشنهادی بهتری تنظیم می‌شود.

در آزمایش‌های انجام شده در مرجع [۱]، تابع خطی تصحیح کننده[۱] به عنوان تابع فعال‌ساز به صورت رابطه (۲-۳۰) استفاده شد.

$$g(a) = \max(0, a) = [\max(0, a_1), \dots, max(0, a_H)] \quad (۳۰-۲)$$

در مرجع [۱] ادعا شد این تابع اغلب عملکرد بهتری از تابع‌های فعال‌ساز دیگر دارد و همچنین عملکرد بهتری را بر روی داده‌های تصویر از خود نشان داده است. از آنجایی که، این تابع یک تابع خطی مبتنی بر قطعه[۲] (زیر) گرادیان با توجه به ورودی آن می‌باشد. این نیاز وجود دارد که پارامترهای گرادیان توسط پس‌انتشار محاسبه شوند که این کار به سادگی توسط رابطه (۲-۳۱) انجام می‌شود:

$$1_{g(a)>0} = [1_{g(a_1)>0}, \dots, 1_{g(a_H)>0}] \quad (۳۱-۲)$$

در اینجا 1_P درصورتی‌که P درست باشد برابر با ۱ می‌باشد در غیر این صورت صفر می‌باشد. از الگوریتم‌های ۱ و ۲ در شکل ۲-۳۲ و شکل ۲-۳۳، برای محاسبه مؤثر توزیع مشترک $p(v, y)$ و گرادیان‌های پارامتر به دست آمده از رابطه (۲-۲۹) برای آموزش نزولی گرادیان تصادفی استفاده شد.

[۱] Rectified

[۲] Piece-Wise

Algorithm 1 Computing $p(\mathbf{v}, y)$ using SupDocNADE

Input: bag of words representation $\mathbf{v}$, target y
Output: $p(\mathbf{v}, y)$
$\mathbf{a} \leftarrow \mathbf{c}$
$p(\mathbf{v}) \leftarrow 1$
for i from 1 to D **do**
  $\mathbf{h}_i \leftarrow \mathbf{g}(\mathbf{a})$
  $p(v_i|\mathbf{v}_{<i}) = 1$
  for m from 1 to $|\pi(v_i)|$ **do**
    $p(v_i|\mathbf{v}_{<i}) \leftarrow p(v_i|\mathbf{v}_{<i})\, p(\pi(v_i)_m|\mathbf{v}_{<i})$
  end for
  $p(\mathbf{v}) \leftarrow p(\mathbf{v})\, p(v_i|\mathbf{v}_{<i})$
  $\mathbf{a} \leftarrow \mathbf{a} + \mathbf{W}_{:,v_i}$
end for
$\mathbf{h}^c(\mathbf{v}) \leftarrow \max(0, \mathbf{a})$
$p(y|\mathbf{v}) \leftarrow \text{softmax}(\mathbf{d} + \mathbf{U}\mathbf{h}^c(\mathbf{v}))_{|y}$
$p(\mathbf{v}, y) \leftarrow p(\mathbf{v})\, p(y|\mathbf{v})$

شکل ۲-۳۲ محاسبه $p(v, y)$ با SupDocNADE

Algorithm 2 Computing SupDocNADE training gradients

Input: training vector $\mathbf{v}$, target y,
unsupervised learning weight λ
Output: gradients of Equation 10 w.r.t. parameters
$f(\mathbf{v}) \leftarrow \text{softmax}(\mathbf{d} + \mathbf{U}\mathbf{h}^c(\mathbf{v}))$
$\delta\mathbf{d} \leftarrow (f(\mathbf{v}) - 1_y)$
$\delta\mathbf{a} \leftarrow (\mathbf{U}^\intercal \delta\mathbf{d}) \circ 1_{\mathbf{h}_y > 0}$
$\delta\mathbf{U} \leftarrow \delta\mathbf{d}\, \mathbf{h}^{c\intercal}$
$\delta\mathbf{c} \leftarrow 0, \delta\mathbf{b} \leftarrow 0, \delta\mathbf{V} \leftarrow 0, \delta\mathbf{W} \leftarrow 0$
for i from D to 1 **do**
  $\delta\mathbf{h}_i \leftarrow 0$
  for m from 1 to $|\pi(v_i)|$ **do**
    $\delta t \leftarrow \lambda(p(\pi(v_i)_m|\mathbf{v}_{<i}) - \pi(v_i)_m)$
    $\delta b_{l(v_i)_m} \leftarrow \delta b_{l(v_i)_m} + \delta t$
    $\delta\mathbf{V}_{l(v_i)_m,:} \leftarrow \delta\mathbf{V}_{l(v_i)_m,:} + \delta t\, \mathbf{h}_i^\intercal$
    $\delta\mathbf{h}_i \leftarrow \delta\mathbf{h}_i + \delta t\, \mathbf{V}_{l(v_i)_m,:}^\intercal$
  end for
  $\delta\mathbf{a} \leftarrow \delta\mathbf{a} + \delta\mathbf{h}_i \circ 1_{\mathbf{h}_i > 0}$
  $\delta\mathbf{c} \leftarrow \delta\mathbf{c} + \delta\mathbf{h}_i \circ 1_{\mathbf{h}_i > 0}$
  $\delta\mathbf{W}_{:,v_i} \leftarrow \delta\mathbf{W}_{:,v_i} + \delta\mathbf{a}$
end for

شکل ۲-۳۳ محاسبه گرادیان‌های آموزش دادن SupDocNADE

۱) نواحی متعدد

اطلاعات مکانی در درک تصویر نقش مهمی را ایفا می‌کند. برای مثال، آسمان اغلب در قسمت بالای تصویر ظاهر می‌شود در حالی که ماشین معمولاً در پایین ظاهر می‌شود. بسیار از کارهای گذشته از این نگرش[۱] به درستی استفاده کرده‌اند. برای نمونه، در مرجع [۸] نشان داده شد که استخراج کلمات بصری متفاوت در مناطق مشخص به جای استفاده از هیستوگرام تصویر، می‌تواند دستاورد قابل‌توجهی در بهبود عملکرد داشته باشد.

در مرجع [۸] مانند روش مشابه، حضور کلمات بصری و شناسایی ناحیه‌ای که ظاهر شده مدل می‌شود. فرض شد که تصویر به چندین ناحیه جدا $R = \{R_1, \dots, R_M\}$ تقسیم شده است که در اینجا M تعداد نواحی می‌باشد. حالا می‌توان تصویر را به صورت رابطه (۲-۳۲) نمایش داد:

$$V^R = [v_1^R, \dots, v_D^R] = [(v_1, r_1), \dots, (v_D, r_D)] \qquad (۲-۳۲)$$

در اینجا، $r_i \epsilon R$ نشان‌دهنده نواحی که کلمه بصری v_i در آن استخراج شده است. از $p(v^R) = \prod_{i=1} p\left((v_i, r_i)|v_{<i}^R\right)$ برای مدل کردن توزیع مشترک تحت کلمات بصری استفاده شد و هر K*M

[۱] Intuition

جفت کلمه/ناحیه ممکن را به‌عنوان یک کلمه مجزا در نظر گرفته می‌شود. یکی از دلایل این کار، رعایت شرط بزرگتر بودن درخت باینری از کلمات بصری می‌باشد تا یک برگ برای هر جفت ناحیه/کلمه بصری ممکن وجود داشته باشد. خوشبختانه، از آنجایی که سرعت رشد محاسبات به صورت لگاریتمی بر اساس اندازه درخت رشد می‌کند مشکل چندانی نیست و هنوز می‌توان تعداد نواحی زیادی را در نظر گرفت.

۲) برچسب‌گذاری

عمل برچسب‌گذاری در تصویر، شامل لیستی از کلمات توصیف کننده محتوای تصویر می‌باشد. برای نمونه، در شکل ۲-۳۱، برچسب‌ها شامل "درخت‌ها" یا "مردم" می‌باشد (برچسب زدن شامل عبارت چندین کلمه به عنوان مثال پیاده‌روی فرد می‌باشد اما آن‌ها به‌عنوان یک نشانه منفرد[۱]می‌باشند). از آنجایی‌که برچسب‌زدن و برچسب به صورت مشخص به هم مرتبط می‌باشند به همین دلیل در این روش پیشنهادی، این دو به صورت مشترک در SupDocNADE مدل شده‌اند.

اگر A واژگان از پیش تعیین‌شده از همه‌ی کلمات برچسب‌گذاری شده باشد. برچسب‌گذاری به دست آمده از یک تصویر را به صورت $a = \{a_1, \dots, a_L\} = and \ \ a \in A$ می‌باشد و L نیز تعداد کلمات موجود در برچسب‌گذاری می‌باشد. در واقع تصویر شامل برچسب‌گذاری را می‌توان ترکیبی از مجموعه کلمات بصری و کلمات برچسب‌گذاری شده نوشت:

$$v^R = \lfloor v_1^A, \dots, v_D^A, v_{D+1}^A, \dots, v_{D+L}^A \rfloor = \lfloor v_1^R, \dots, v_D^R, a_1, \dots, a_L \rfloor \qquad (۳۳-۲)$$

باید کلمات برچسب‌گذاری در SupDocNADE تعبیه شوند. در این مقاله از اندیس‌گذاری[۲]مشترک برای کلمات بصری و برچسب آنها استفاده شد. از تعداد زیادی درخت کلمه باینری برای تقویت آن استفاده شده که برگ‌ها به عنوان برچسب کلمات در نظر گرفته شده‌اند. آموزش SupDocNADE بر روی اتصالات تصویر/برچسب‌گذاری، با V^A نمایش داده شد که این می‌تواند رابطه بین برچسب‌ها را در مکان تعبیه‌شده از کلمات بصری و کلمات برچسب‌گذاری شده یاد بگیرد.

[۱] Token
[۲] Indexing

در زمان آزمون، کلمات برچسب‌گذاری نشده است و باید پیش‌بینی شوند. برای رسیدن به این هدف، نمایش اسناد برپایه فقط کلمات بصری محاسبه می‌شود و برای هر کلمه، برچسب‌گذاری ممکن با احتمال این که کلمه بتواند در کلمه بعدی مشاهده شود محاسبه می‌شود. این کار بر پایه تجزیه درخت مانند رابطه (۲-۲۲) انجام می‌شود. به عبارت دیگر، در [۱] فقط احتمال مسیرهایی که به یک برگ متناظر می‌رسد را در کلمه برچسب‌گذاری (نه یک کلمه بصری) محاسبه می‌کند. سپس کلمات بصری در A رتبه‌بندی می‌شوند و پنج کلمه با بالاترین احتمال به عنوان برچسب‌گذاری‌های پیش‌بینی‌شده انتخاب می‌شوند.

مرجع [۲] یکی دیگر از روش‌های موفق در دسته‌بندی صحنه بر پایه مدل‌های موضوعی می‌باشد که بر استخراج ویژگی‌های موضوعی تاکید دارد. در همین راستا در ادامه روش ارائه شده در این مقاله و همچنین روش ارائه شده برای استخراج ویژگی‌های موضوعی به تفصیل بیان می‌شود.

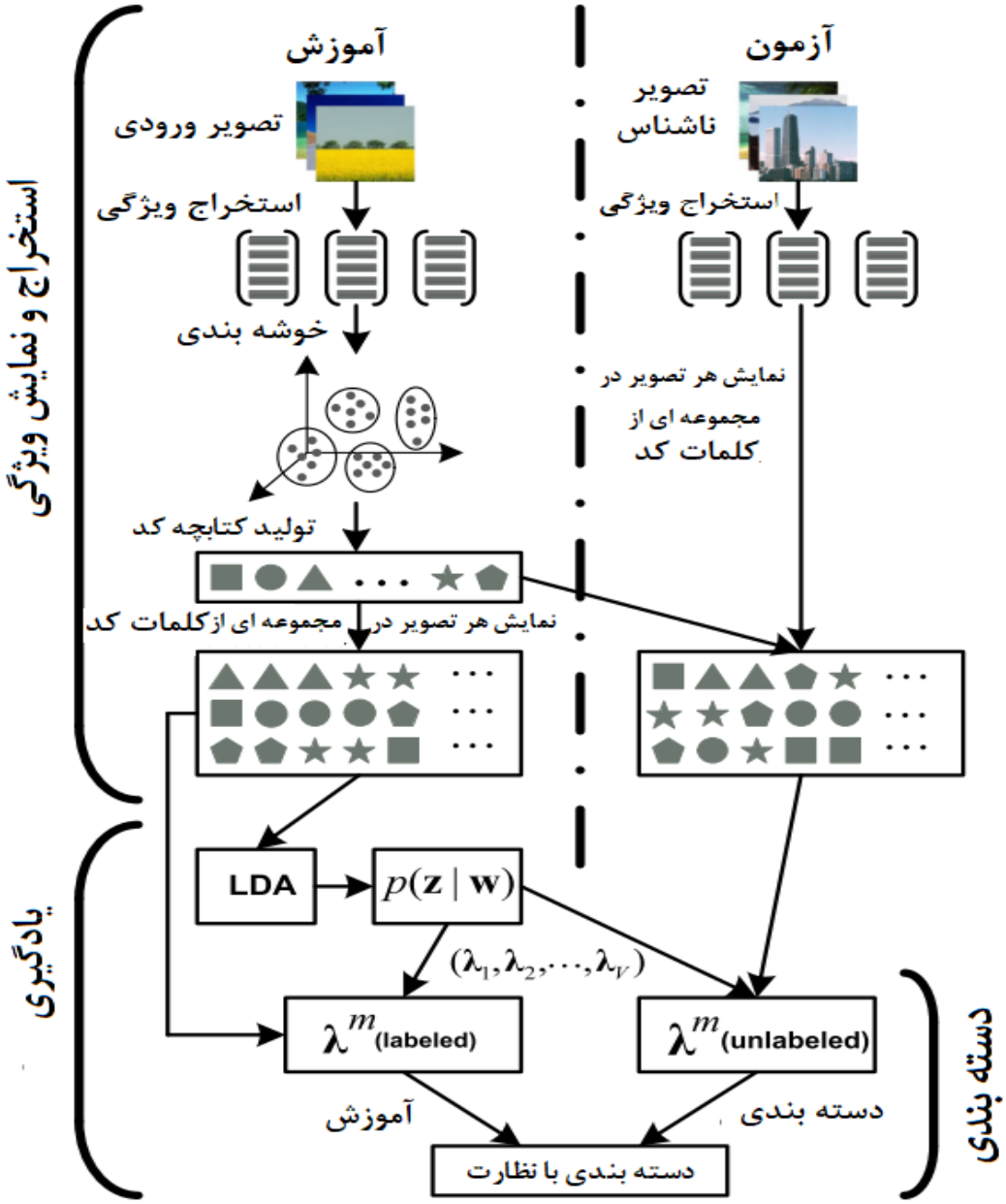
۲ - ۱۲ - ۲ استخراج ویژگی موضوعی برای دسته‌بندی

مرجع [۲] یکی از کارهای موفق انجام‌شده در سال‌های اخیر می‌باشد. در این مرجع یک ویژگی موضوعی برای دسته‌بندی صحنه ارائه شد. این ویژگی بر پایه نمایش روش‌های موضوعی[۱] از تصویر ساخته‌شده است. تفاوت ویژگی معرفی شده در این کار نسبت به کارهای دیگر، معرفی ویژگی جدید که موضوع‌ها را در کلاس‌های متفاوت به اشتراک می‌گذارد و نیاز به برچسب کلاس‌ها قبل از دسته‌بندی نمی‌باشد. برای نمایش یک تصویر جدید، روش پیشنهادی به صورت مستقیم ویژگی موضوع را با استفاده از نگاشت خطی کلمه‌های کد[۲] به‌جای استنتاج متغیرهای پنهان استخراج می‌کند. در ادامه به تشریح این روش پرداخته می‌شود.

---

[۱] Thematic

[۲] Codewords



شکل ۲-۳۴ روش پیشنهادی دسته‌بندی صحنه بر پایه ویژگی‌ها موضوعی بر اساس مرجع [۲]

در الگوریتم مرجع [۲] که در شکل ۲-۳۴ نشان داده شد، یک تصویر به عنوان مجموعه‌ای از وصله‌های محلی در نظر گرفته می‌شود و کلمه‌های کد تصویر توسط خوشه‌بندی وصله‌ها تولید می‌شوند. کلمه‌های کد به عنوان مراکز خوشه‌های آموزش داده شده معرفی می‌شوند. یک مجموعه کامل از کلمه‌های کد به عنوان کتابچه کد[۱] در نظر گرفته می‌شوند. هر تصویر به عنوان مجموعه‌ای از کلمه‌های کد توسط

[۱] Codebook

وصله‌هایش نمایش داده می‌شود. در مجموعه آموزش، یک مدل LDA از طریق کتابچه کد ساخته می‌شود و سپس از آن برای تولید فضای ویژگی‌های موضوعی استفاده می‌گردد. کلمه‌های کد هر تصویر توسط روش نگاشت خطی پیشنهادی به بردارهای ویژگی تحت فضای ویژگی موضوعی تبدیل می‌شوند. بردارهای ویژگی برچسب‌گذاری شده تصاویر برای آموزش دسته‌بندی بانظارت استفاده می‌شوند. در مرحله آزمون، یک تصویر در ابتدا، به وسیله همان کتابچه کد نمایش داده می‌شود و سپس بردارهای ویژگی مرتبط در فضای ویژگی موضوعی مشابه تولید می‌شود. در نهایت تصویر مجموعه آزمون به وسیله دسته‌بند آموزش داده شده و از طریق بردارهای ویژگی آنها دسته‌بندی می‌شوند.

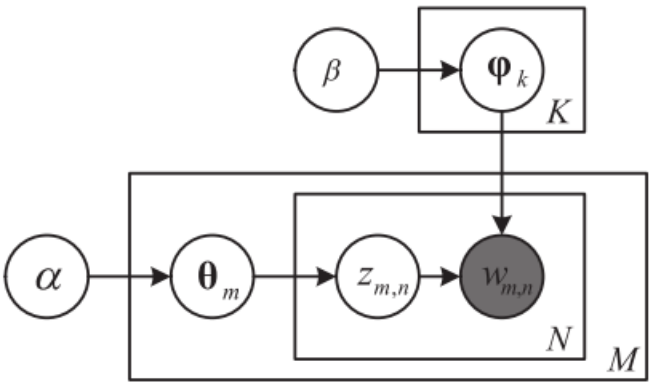
۱) ویژگی و کتابچه کد

بسیاری از مدل‌های غیراحتمالی[۱]برای دسته‌بندی صحنه طبیعی بیشتر بر ویژگی‌های سراسری مانند توزیع تکرار[۲](فرکانس)، جهت‌های[۳]لبه و هیستوگرام رنگ تمرکز دارند و مدل‌های موضوعی احتمالی از ویژگی‌های محلی برای دسته‌بندی استفاده می‌کنند. ویژگی‌های محلی در برابر انسدادها[۴]و تغییراتی مکانی[۵]نسبت به ویژگی‌های سراسری مقاومتر[۶] هستند، در نتیجه انتظار می‌رود روشی برای نمایش نواحی محلی تصویر ارائه کرد. بنابراین چهار راه متفاوت برای استخراج نواحی محلی در مرجع [۷۱] معرفی شد. در این مرجع نشان داده شد که توصیفگر ناحیه ۱۲۸بعدی SIFT اطلاعات بهتر و قویتری را فراهم می‌کند. این ویژگی‌های محلی در دامنه وسیعی از دسته‌بندی صحنه مورد استفاده قرار گرفته‌اند. در مرجع [۲] یک فرم متفاوت از SIFT به کار گرفته شد که در آن از شبکه منظم متراکم به جای نقاط مورد علاقه استفاده شده است. این روش، رویه نرمال‌سازی SIFT معمولی را وقتی‌که بزرگی شیب سراسری متعلق به مسیر حرکت بسیار ضعیف است نادیده می‌گیرد. برای محاسبه توصیفگر ناحیه در هر وصله، در گام نخست یک تصویر به وسیله شبکه مشبک لغزان به وصله‌های دارای همپوشانی تقسیم

---

[۱] Non-Probabilistic
[۲] Frequency
[۳] Orientations
[۴] Occlusions
[۵] Spatial Variations
[۶] Robust

می‌شود. در گام بعدی از الگوریتم k-means برای خوشه‌بندی توصیفگر ناحیه SIFT استفاده شد و کلمه‌های کد به عنوان مراکز خوشه‌ها تعریف می‌شوند که همه‌ی آن‌ها به عنوان کتابچه کد معرفی می‌شوند.



شکل ۲-۳۵ مدل گرافیکی از نمایش LDA

۲) ساختار مدل

شکل ۲-۳۵ نمایشی از مدل گرافیکی LDA برای پردازش تصویر زا نمایش می‌دهد. در این مدل φ_k، احتمال کلمه کد در موضوع k را نشان می‌دهد، θ_m نمایش احتمال موضوع در تصویر mام می‌باشد. θ_m و φ_k به ترتیب برای تولید موضوعات و کلمه‌ها به عنوان پارامترهای توزیع چند جمله‌ای استفاده شده است. k تعداد موضوعات، M تعداد تصاویر پایگاه داده، N تعداد کلمات کد در هر تصویر می‌باشند. $w_{m;n}$ و $z_{m;n}$ به ترتیب نشان‌دهنده nامین کلمه کد و موضوع آن در mامین تصویر می‌باشد. پارامتر α و β پارامترهای توزیع دیریکله می‌باشند. در [۲] از مدل گرافیکی مشابه شکل ۴(a) متعلق به مرجع [۸۵] که برای نمایش فرآیند یادگیری و تولید معرفی شد، استفاده می‌کند. این روش اساساً مانند روش LDA اصلاح شده در تصویر از مرجع [۶۹] می‌باشد که β_k اصلی، به عنوان φ_k نمایش داده می‌شود. در واقع باید اشاره نمود که LDA بر پایه تکنیک‌های پردازش متن برای پردازش تصویر به کار گرفته شد که در ادامه شباهت بین اصطلاحات تصویر و متن بیان می‌شود:

- کلمه کد w، یک واحد پایه‌ای از تصویر می‌باشد که برای عضویت کلمه کد از ایندکس دیکشنری کلمه‌های کد $\{1, \dots, v\}$ استفاده شده است. کلمه کد vام، در دیکشنری توسط w بردار-v

نمایش داده شد به طوری که برای $v \neq t$ داشته باشیم $w^v = 1$ و $w^t = 0$. در تصویر بالا، w سایه‌دار شده که نشان دهنده یک متغیر قابل مشاهده می‌باشد. یک کلمه کد معادل یک "کلمه" در پردازش متن می‌باشد.

- یک تصویر، توالی از N وصله می‌باشد که توسط $w_m = (w_1, \dots, w_N)$ $(m = 1,2, \dots, M)$ نشان داده می‌شود، w_n، nامین وصله از تصویر می‌باشد. هر تصویر معادل یک "سند"در پردازش متن می‌باشد.
- $W = (w_1, \dots, w_m)$ نمایش دهنده تصاویر پایگاه داده می‌باشند و این معادل با "مجموعه داده متنی" در پردازش متن می‌باشد.

حال می‌توان روند کلی تولید یک تصور w_n از مدل LDA را به صورت زیر بیان نمود:

۱. نمونه‌برداری از موضوع توسط $\varphi_k \sim Dirichlet(\beta), k \in [1, k]$ انجام می‌شود.

۲. برای یک تصویر (w_n) در پایگاه داده w، نمونه‌برداری توزیع احتمالی موضوع توسط $\theta_m \sim Dirichlet(\alpha)$ انجام می‌شود.

۳. برای nامین کلمه کد ($w_{m,n} \in [1, N]$) در تصویر مراحل زیر انجام می‌شود:

a. انتخاب موضوع پنهان توسط $z_{m,n} \sim Multinomial(\theta_m)$ صورت می‌گیرد.

b. تولید یک کلمه کد به وسیله $w_{m,n} \sim Multinomial(\varphi_{z_{m,n}})$ انجام می‌شود.

θ_m و φ_k توزیع دیریکله می‌باشد که توزیع پیشین مزدوج[۲] از توزیع چند جمله‌ای می‌باشند. تابع توزیع به صورت رابطه (۲-۳۴) طراحی می‌شود:

$$Dir(\mu|\alpha) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1), \dots, \Gamma(\alpha_k)} \prod_{k=1}^{K} \mu_k^{\alpha_k - 1} \qquad (۲-۳۴)$$

[۱] Corpus

[۲] Conjugate

در اینجا، $0 \le \mu_k \le 1, \sum_k \mu_k = 1, \alpha_0 = \sum_{k=1}^{K} \alpha_k$ و $\Gamma(.)$ تابع گاما استاندارد می‌باشد. با توجه به ابرپارامترهای α و β، توزیع مشترک از متغیرهای پنهان و مشخص می‌تواند به صورت رابطه های (۲-۳۵)، (۲-۳۶) و (۲-۳۷) تعریف شود:

$$p(w_m, z_m, \theta_m, \phi|\alpha, \beta) = \prod_{n=1}^{N} p(w_{m,n}|\varphi_{z_{m,n}}) P(z_{m,n}|\theta_m).\, p(\theta_m|\alpha).\, p(\phi|\beta) \quad (۲-۳۵)$$

$$p(w_{m,n} = t, \theta_m, \phi) = \sum_{k=1}^{K} p(w_{m,n} = t|\varphi_k) p(z_{m,n} = k|\theta_m) \quad (۲-۳۶)$$

$$p(w|\theta, \Theta) = \prod_{m=1}^{M} p(w_m|\theta_m, \phi) = \prod_{m=1}^{M} \prod_{n=1}^{N} p(w_{m,n}|\theta_m, \phi) \quad (۲-۳۷)$$

در اینجا، $\phi = \{\varphi_k\}_{k=1}^{K}$ یک ماتریس با اندازه $K \times V$، $\Theta = \{\theta_m\}_{m=1}^{M}$ یک ماتریس با اندازه $M \times K$ می‌باشد.

۳) تخمین پارامتر، نمونه‌بردار گیبس[۱]

برای تخمین پارامترهای مدل LDA روش‌هایی همچون تقریب لاپلاس ، استنتاج تغییرات[۲] و زنجیره مارکوف مونت‌کارلو[۳](MCMC) وجود دارند [۸۶]. نمونه‌بردار گیبس یک نوع خاص از MCMC می‌باشد که یک جزء از توزیع مشترک را از دیگر متغیرهای هر مرحله که ثابت هستند نمونه می‌گیرد. نمونه‌بردار گیبس می‌تواند الگوریتم نسبتاً ساده‌ای تولید کند که دارای بُعد توزیع مشترک بالا می‌باشد. هر روش تخمین‌زننده دارای مزایا و معایبی می‌باشد و فاکتورهای زیادی مانند بهره‌وری، پیچیدگی، دقت و راحتی محتوا برای انتخاب روش مناسب برای تخمین الگوریتم استنتاج نیاز می‌باشد. از آنجایی‌که نمونه‌بردار گیبس برای توصیف و پیاده‌سازی ساده می‌باشد این روش برای تخمین پارامترهای مورد استفاده قرار می‌گیرد.

هدف از این روش محاسبه توزیع پسین در رابطه (۲-۳۸) می‌باشد:

---

[۱] Gibbs

[۲] Variational Inference

[۳] Markov chain Monte Carlo(MCMC)

$$p(Z|W) = \frac{p(W,Z)}{\sum_z p(W,z)} \quad (۳۸-۲)$$

این توزیع به صورت مستقیم قابل محاسبه نمی‌باشد چون مجموع در مخرج نمی‌تواند فاکتورایز شود و تعداد ترم‌های آن بسیار زیاد می‌باشد که k^n ترم وجود دارد (n در اینجا مجموع تعداد نمونه‌های کلمه‌های کد در تصاویر پایگاه داده می‌باشد). در همین راستا می‌توان از نمونه‌برداری گیبس با استخراج نمونه‌های تنها با متغیر پنهان (موضوع) برای برطرف کردن مشکل زمانبر بودن، استفاده کرد. به صورت خاص نمونه‌بردار گیبس برای این مدل، موضوع z را از هر کلمه کد w، که از تخمین‌زدن واقعی پارامترهای θ_m و φ_k اجتناب می‌کند، نمونه‌برداری می‌کند. در این روش موضوع هر کلمه کد مشخص می‌شود و مقادیر θ_m و φ_k می‌توانند از آمار تکرار[۱] محاسبه شوند. در نهایت روابط نمونه‌برداری به‌صورت رابطه (۳۹-۲) بیان می‌شود [۶۹]:

$$\begin{aligned} p(z_i = k|z_{-i}, w) &\propto \frac{n_{k,-i}^{(t)} + \beta_t}{\sum_{v=1}^{V} n_{k,-i}^{(v)} + \beta_v} \cdot \frac{n_{m,-i}^{(k)} + \alpha_k}{[\sum_{z=1}^{k} n_m^{(z)} + \alpha_z] - 1} \\ &\propto \frac{n_{k,-i}^{(t)} + \beta_t}{\sum_{v=1}^{V} n_{k,-i}^{(v)} + \beta_v} \left(n_{m,-i}^{(k)} + \alpha_k\right) \end{aligned} \quad (۳۹-۲)$$

در اینجا فرض شده که $w_i = t$ نشان‌دهنده iامین کلمه کد متناظر با متغیر موضوع، تعداد $n_{.,-i}^{(.)}$ بیانگر حذف نشانه[۲] i از موضوع متناظر است. $n_k^{(v)}$ تعداد دفعه‌هایی که کلمه کد v به موضوع k انتساب داده شده است، β_v دیریکله پیشین از کلمه کد v، $n_m^{(z)}$ تعداد دفعاتی که موضوع z به تصویر m انتساب داده شده و α_z دیریکله پیشین از موضوع z می‌باشد. بعد از نمونه‌برداری می‌توان توسط رابطه‌های (۴۰-۲) و (۴۱-۲)، مقادیر φ و θ را از مقدار z تخمین زد:

$$\varphi_{k,t} = \frac{n_k^{(t)} + \beta_t}{\sum_{v=1}^{V} n_k^{(v)} + \beta_v} \quad (۴۰-۲)$$

$$\theta_{m,k} = \frac{n_m^{(k)} + \alpha_t}{\sum_{z=1}^{Z} n_m^{(z)} + \alpha_z} \quad (۴۱-۲)$$

[۱] Frequency Statistics

[۲] Token

در این رابطه، $\varphi_{k,t}$ احتمال انتساب کلمه کد t به موضوع k و $\theta_{m,k}$ احتمال انتساب موضوع k به تصویر m می‌باشند. با استفاده از معادلات (۲-۳۹)، (۲-۴۰) و (۲-۴۱) می‌توان مراحل نمونه‌برداری گیبس را به آسانی انجام داد (برای بررسی جزئیات بیشتر می‌توان به مرجع [۶۹] رجوع کرد).

۴) نمایش تصویر با ویژگی‌های موضوعی

همان‌طور که در مراحل پیشین اشاره شد، تصاویر به عنوان یک مجموعه‌ای از کلمه‌های کد مدل می‌شوند. وظیفه دسته‌بندی صحنه تصویر با استفاده از مدل موضوعی مشابه زبان طبیعی بهبود می‌یابد. در همین راستا، روش LDA برای زبان طبیعی در تصویر اعمال شد. به دست آوردن یک نمایش از تصویر جدید در فضای موضوع بسیار راحت می‌باشد. در تصاویر داده شده، کلمه‌های کد، مدل آموزش داده شده (Mod) و موضوع پنهان در هر کلمه کد می‌تواند توسط رابطه (۲-۴۲) نمونه‌برداری شود:

$$p(\widetilde{z_i} = k|\widetilde{w_i} = t, \widetilde{z_{-i}}, \widetilde{w_{-i}}; Mod) \propto \varphi_{k,t}(n_{\tilde{m},-i}^{(k)} + \alpha_k) \qquad (۲-۴۲)$$

$\tilde{z}$موضوع تصویر جدید $\tilde{w}$ می‌باشد، موضوع‌ها و آمار آن‌ها، توسط تکرار در نمونه‌برداری گیبس بروز می‌شوند. احتمال پسین موضوع‌ها می‌تواند توسط رابطه (۲-۴۲) به دست بیاید. بعد از نمونه‌برداری، رابطه (۲-۴۱) توزیع موضوع را برای تصویر جدید به صورت رابطه (۲-۴۳) محاسبه می‌کند:

$$\theta_{m,k} = \frac{n_{\tilde{m}}^{(k)} + \alpha_k}{\sum_{z=1}^{K} n_{\tilde{m}}^{(z)} + \alpha_z} \qquad (۲-۴۳)$$

این فرآیند برای مجموعه کاملی از تصاویر جدید قابل استفاده می‌باشد [۶۹]. در صورتی که یک فرض ساده انجام شود که برای هر تصویر جدید، $\varphi_{k,t}$ بروزرسانی نشده و $\theta_{\tilde{m},k}$ برابر با $E(\theta_{m,k})$ می‌باشد، احتمال پسین برای تصویر جدید می‌تواند به سادگی توسط رابطه (۲-۴۴) محاسبه شود:

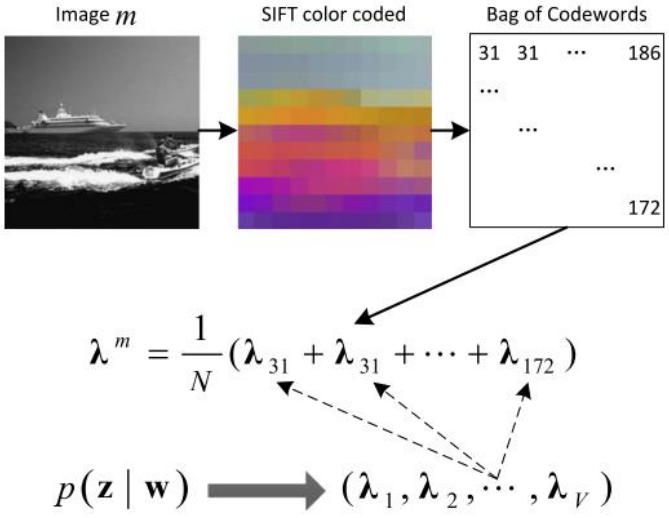
$$p(\tilde{z} = k|\tilde{w}; Mod) \propto \frac{1}{M}\sum_{m=1}^{M} \theta_{m,k} \cdot \varphi_{k,t} \qquad (۲-۴۴)$$

M تعداد تصاویر در مجموعه آموزش می‌باشند. توسط این عملگر، تمامی تصاویر در یک فضای موضوعی یکسان نمایش داده می‌شوند. روش نمایش دادن یک تصویر توسط مدل LDA، معادل نگاشت تصاویر

در فضایی می‌باشد که در آن، بردار $p(z|w_j)$ به فرم پایه خودش باشد و کلمات کد مانند مختصات[۱] باشند. این فضای ویژگی متحد[۲]به استفاده از این نوع نمایش موضوع، به عنوان یک ویژگی برای دسته‌بندی بانظارت می‌باشد. در این روش، زمانی که یک تصویر جدید نمایش داده می‌شود، نیازی به ایجاد $\theta_{\tilde{m}}$ و $n_{\tilde{m},-i}^{(k)}$ نمی‌باشد. بنابراین گام‌های مربوط به رابطه‌های (۲-۴۲) و (۲-۴۳) می‌توانند در مجموع حذف شوند. احتمال پیشین $p(z|w)$ را می‌توان به وسیله ماتریس $k \times V$ بعدی A توصیف نمود که هر عنصر A_{ij} احتمال انتساب jامین کلمه کد به iامین موضوع می‌باشد. در اینجا $\lambda_j A(:,j)$ به گونه‌ای تعریف می‌شود که یک تصویر معادل مجموعه‌ای از کلمات کد می‌باشد. بنابراین یک وصله از یک تصویر به وسیله کلمه کد λ_j خودش نمایش داده می‌شود. یک تصویر m با کلمات کد $w_m = (w_1 + \cdots + w_N)$ می‌تواند به عنوان ویژگی‌های موضوعی به صورت رابطه (۲-۴۵) تعریف شود:

$$\lambda^m = \frac{1}{N}(\lambda_{w_1} + \cdots + \lambda_{w_N}) \qquad (۴۵-۲)$$

شکل ۲-۳۶ رابطه بین تصویر m با λ^m را نشان می‌دهد.



شکل ۲-۳۶ ویژگی‌های موضوعی استخراج شده از یک تصویر بر اساس مرجع [۲]

[۱] Coordinates
[۲] Unified

نمایش یک تصویر شامل دو قسمت، کلمات کد یک تصویر و اطلاعات آماری از پایگاه داده تصویر می‌باشد. باید به این نکته اشاره کرد که استخراج این ویژگی از یک تصویر جدید یک نگاشت تحت کلمات کد می‌باشد که زمان پردازش به صورت چشمگیری کاهش می‌یابد.

۲ - ۱۳ جمع بندی

در این فصل کارهای انجام شده برای دسته‌بندی صحنه به تفصیل بیان شد. یک فاز مهم در این نوع از سیستم‌ها به دست آوردن نواحی کاندیدا می‌باشد. در همین راستا در این فصل، روش‌های استخراج نواحی کاندیدا و نقاط قوت و ضعف هر یک از این روش‌ها معرفی شد. با توجه به مطالب ارائه شده در این فصل، روش‌های استخراج نواحی کاندیدا براساس یادگیری عمیق بهترین عملکرد را در صحنه های پیچیده از خود نشان داده است. در ادامه این فصل، کارهای انجام شده در دسته‌بندی صحنه‌هایی با پیچیدگی کم مانند تصاویر نامتعارف مورد بررسی قرار گرفته است. مطالب بیان شده در این فصل نشان می‌دهد که استفاده از روش‌های موجود برپایه ویژگی‌های سطح پایین نمی‌توانند در این نوع از محیط‌ها عملکرد مناسبی از خود نشان دهند. از آنجایی که معمولاً عملکرد این نوع از سیستم‌ها در سامانه‌های تحت وب مورد ارزیابی قرار می‌گیرد زمان پردازش آنها بسیار مهم می‌باشد. در ادامه این فصل کارهای انجام شده در دسته‌بندی صحنه‌های پیچیده معرفی شد. در سال‌های اخیر معمولاً از ویژگی‌های سطح پایین مانند رنگ و بافت برای دسته‌بندی استفاده شد ولی با افزایش تعداد کلاس‌ها و بالا رفتن پیچیدگی اشیاء موجود در تصویر، این نوع از ویژگی‌ها به تنهایی نمی‌توانند عملکرد خوبی را در دسته‌بندی صحنه از خود نشان دهند. در تحقیقات انجام شده در سال‌های اخیر، روش‌ها برپایه ویژگی‌های سطح بالا مانند مدل‌های موضوعی برای دسته‌بندی صحنه استفاده شده است. در این فصل روش‌های موجود در این حوزه به تفصیل بیان شد و نقاط قوت و ضعف هر روش مورد بررسی قرار گرفت. با توجه به مطالب ارائه شده در این فصل، به دست آوردن نواحی کاندیدا و برچسب این نواحی و موضوعات موجود در تصویر

می‌تواند در بالا بردن دقت سیستم دسته‌بندی تصویر بسیار موثر باشد. در همین راستا در فصل بعدی روشی جامع بر اساس این روش‌ها معرفی می‌شود

۳ روش پیشنهادی در دسته‌بندی موضوعی تصاویر

۳ - ۱ مقدمه

در این فصل روش پیشنهادی به منظور دسته‌بندی صحنه تصویر ارائه می‌شود. همانطور که قبلا اشاره شد، در این رساله دو دیدگاه برای دسته‌بندی تصاویر بر اساس تعداد کلاس درنظر گرفته شد. در دیدگاه نخست، تصاویری با دو دسته مورد بررسی قرار می‌گیرند. برای مثال تصاویر فضای آموزشی/غیرآموزشی، ورزشی/غیردرسی و متعارف/غیرمتعارف که در آن یک دسته شامل موضوع خاصی می‌باشد و دسته دیگر می تواند موضوعات متنوعی را داشته باشد در واقع تشخیص موضوع خاص در این مسئله مهم می‌باشد. در این رساله، با توجه به اهمیت دسته‌بندی تصاویر نامتعارف روشی به منظور دسته‌بندی این نوع از تصاویر معرفی شده است. در واقع در این دسته‌بندی تصاویر شامل خصیصه‌های مرتبط با موضوع نامتعارف مهم می‌باشد و خصیصه‌های نامتعارف می‌تواند در هر موضوعی باشد. در ادامه این روش به تفصیل بیان می‌شود.

در دیدگاه دوم روشی برای دسته‌بندی تصاویر با تعداد دسته‌های زیاد و پیچیدگی بالا پیشنهاد می‌شود. روش پیشنهادی بر اساس یادگیری عمیق و یک دیدگاه نوین برای استخراج مدل موضوعی و ویژگی استخراج شده از نواحی مرتبط به اشیاء در تصویر می‌باشد. جزئیات این روش در بخش‌های بعدی بیان می‌شود.

۳ - ۲ روش پیشنهادی برای دسته‌بندی تصاویر نامتعارف

امروزه با پیشرفت اینترنت و رسانه‌های تحت وب، توزیع و اشتراک منابع اطلاعاتی مانند تصویر در حال افزایش است. اشتراک منابع، خطرات و مشکلاتی نظیر دسترسی افراد پایین‌تر از سن قانونی به تصاویر نامتعارف را ایجاد می‌کند. در این بخش روشی نوین برای طراحی سیستمی هوشمند به منظور تحلیل و فیلترینگ تصاویر تحت وب معرفی می‌شود که در حجم بالای داده‌ها نیز کاربردی است. روش ارائه شده بر پایه شبکه‌های عصبی پیچش با معماری نوین و مقاوم نسبت به حجم بالای داده می‌باشد [۸۷].

۳ - ۲ - ۱ شبکه‌های عصبی عمیق

امروزه شبکه‌های عصبی به دلیل قابلیت مناسب در یادگیری ویژگی‌های ترکیبی و نوین، در دسته‌بندی استفاده می‌شوند. شبکه‌ی عصبی عمیق، مدلی از شبکه عصبی است که برای یادگیری تبدیل غیرخطی روی داده‌ها مناسب است. تفاوت اساسی روش‌های شبکه عمیق با شبکه عصبی معمولی، حفظ قابلیت بازسازی داده در هر لایه است. به عبارت دیگر شبکه‌ی عمیق، با مدل کردن توزیع داده‌ها در هر لایه، اطلاعات فضای ویژگی را حفظ می‌کند. این شبکه، قابلیت تعمیم‌پذیری بیشتری روی داده‌های آزمون دارد و در آن بیش‌برازش[۱]مدل رخ نمی‌دهد. در واقع در شبکه‌های عصبی معمولی وجود لایه‌هایی با اتصال کامل سبب اتلاف[۲]می‌شود و از طرفی تعداد زیاد پارامتر در این نوع از شبکه‌ها، سرعت بیش‌برازش را افزایش می‌دهند. ولی در شبکه‌های عصبی عمیق به دلیل استفاده از تکنیک‌های اشتراک وزن،[۳]ناحیه ادراکی محلی[۴]و نمونه‌گیری مکانی کاهشی[۵]احتمال وقوع بیش‌برازش کاهش می‌یابد. در برخی از کاربردها، استخراج ویژگی‌های معنادار دشوار است و یا ویژگی‌های معنادار استخراج شده، بازنمایی مناسبی از داده‌ها را ندارند. در این راستا، استخراج ویژگی‌ها از طریق یادگیری دارای اهمیت می‌باشد. یادگیری عمیق، بر اساس یادگیری چند سطحی از بازنمایی‌های مختلف بر روی یک ساختار سلسه‌مراتبی از ویژگی‌ها است. در یادگیری عمیق، مفاهیم سطح بالا از روی ویژگی‌های سطح پایین تعریف می‌شوند و مفاهیم سطح پایین نیز به تعریف مفاهیم سطح بالاتر کمک می‌کنند [۸۸]. هدف یادگیری عمیق، استخراج هوشمندانه ویژگی‌ها طی یک مرحله یادگیری است که این موضوع یکی از تحقیقات مهم تیم گوگل در سال ۲۰۱۲ بود که برای تشخیص مفاهیم سطح بالاتر انجام شد [۸۹].

روش‌های بسیاری برای یادگیری شبکه عمیق عمیق در سال‌های اخیر معرفی شده است که این روش‌ها را می‌تون در حالت کلی به دسته‌های، یادگیری مستقل از حالت،[۶] یادگیری پویا،[۷] یادگیری بر اساس زیان و گرادیان تقسیم نمود[۹۰]. نتایج بدست آمده در سال‌های اخیر در کاربرد دسته‌بندی تصویر و

[1] Overfitting
[2] Wasteful
[3] Weight Sharing
[4] Local Receptor Field
[5] Spatial Down Sampling
[۶] Independent Mode Learning
[۷] Dynamic

تشخیص اشیاء نشان می‌دهد که یادگیری برپایه گرادیان و تابع زیان بهترین عملکرد را در شبکه‌های عصبی عمیق دارد [۹۱]. در این رساله برای یادگیری شبکه عصبی عمیق از این روش استفاده شده است.

در روش پیشنهادی برای دسته‌بندی تصاویر نامتعارف، از شبکه‌های عمیق برای استخراج ویژگی‌های مناسب و طبقه‌بندی تصاویر نامتعارف استفاده گردیده که روشی نوین و کارا در این زمینه است. شبکه‌های عمیق انواع متفاوتی دارند که از مهم‌ترین آن‌ها می‌توان به شبکه‌عصبی پیچش، شبکه باور عمیق[۱] شبکه عصبی حافظه کوتاه مدت ماندگار[۲]و غیره اشاره نمود. یکی از شبکه‌های عمیق استفاده شده در پردازش تصویر، شبکه‌های پیچش است که در این تحقیق نیز، از این نوع شبکه استفاده گردید. در ادامه شبکه پیچش تشریح می‌شود.

۳ - ۲ - ۲ شبکه‌های عصبی پیچش

ورودی‌های شبکه‌های عصبی پیچش، تصاویر هستند. در ساختار این نوع از شبکه‌ها از تکنیک‌های اصلی شامل اشتراک وزن، ناحیه دریافت محلی و نمونه‌گیری مکانی کاهشی استفاده می‌شود. این شبکه تا حدی به مدل سیستم بینایی بیولوژیکی انسان شبیه است [۹۲]. از لحاظ مفهومی، شبکه‌های عصبی پیچش را می‌توان سیستمی از چندین تشخیص‌دهنده سلسله مراتبی متصل به هم در نظر گرفت.

در لایه‌ی اول این شبکه، معمولاً تطبیق الگو[۳]غیرخطی با رزولوشن مکانی مناسب انجام می‌گیرد و ویژگی‌های اصلی داده ورودی استخراج می‌شوند. لایه بعدی وجود ترکیب مکانی ویژگی‌های قبلی را تشخیص می‌دهد. گاهی عمل نمونه‌برداری کاهشی نیز با این شبکه همراه است. در این صورت، لایه‌های بعدی عمل شناسایی الگو را در مقیاس مکانی وسیع‌تر و با رزولوشن پایین‌تر انجام می‌دهند. با استفاده

[۱] Deep Belief Network
[۲] Long Short-Term Memory
[۳] Template Matching

از خاصیت نمونه‌برداری کاهشی در شبکه‌عصبی پیچش، می‌توان با تعداد محدودی وزن، ورودی‌های بزرگی را پردازش نمود.

لایه پیچش (ConV) هسته اصلی تشکیل دهنده شبکه عصبی پیچش است. این لایه شامل نورون‌هایی در قالب توده‌های سه بعدی با عرض، ارتفاع و عمق مشخص می‌باشد. تعداد نورون‌های هر توده، متناسب با فضای تصویر ورودی تعیین می‌شوند. خروجی این لایه، یک توده با توزیع مکانی را تشکیل می‌دهد که توسط لایه پیچش بعدی مورد پردازش قرار می‌گیرد و در نهایت به لایه خروجی منتقل می‌شود. در واقع لایه بعدی توسط لغزاندن فیلتر بر روی توده لایه فعلی به دست می‌آید.

استفاده از نمونه‌برداری کاهشی یا انتخاب در بین لایه‌ها، رزولوشن لایه‌های بعدی را کاهش می‌دهد. لایه انتخاب معمولاً بین چندین لایه پیچش قرار می‌گیرد. این لایه موجب کاهش اندازه تصویر و در نتیجه کاهش محاسبات در داخل شبکه و کنترل بیش‌برازش می‌شود. لایه انتخاب به صورت مستقل بر روی هر برش عمقی از توده ورودی عمل می‌کند و آن‌را از لحاظ مکانی تغییر اندازه می‌دهد. از طرفی باید اشاره کرد که این لایه خاصیت تغییرناپذیری انتقال مکانی[1] دارد. در واقع نسبت به جابجایی مقاوم است. تابع‌های انتخاب بسیاری برای شبکه‌های پیچش مانند تابع ماکزیمم‌گیری و تابع متوسط‌گیری معرفی شده است.

لایه تابع فعال‌ساز معمولاً قبل از لایه انتخاب در این نوع از شبکه‌ها قرار دارد. در این لایه بر روی تک‌تک نورون‌ها یک تابع فعال‌ساز اعمال می‌شود. همان‌طور که اشاره شد در اینجا از تابع فعال‌ساز با عنوان واحد خطی اصلاح‌کننده (ReLU[2]) استفاده شده است. در این لایه بر روی تک‌تک نورون‌ها یک تابع فعال‌ساز اعمال می‌شود. لایه بعدی استفاده شده در این نوع از شبکه‌ها، لایه تمام متصل (FC[3]) می‌باشد. نورون‌هایی که در این لایه قرار دارند با تمام نورون‌های موجود در لایه قبلی (مشابه شبکه‌های عصبی

[1] Translation Invariance
[2] Rectified Linear Unit (ReLU)
[3] Fully Connected (FC)

معمولی) ارتباط دارند. بنابراین می‌توان از ضرب ماتریسی و سپس جمع نتیجه حاصله با بایاس برای محاسبه تمامی نورون‌ها استفاده کرد.

با توجه به توضیحات ارائه شده، در ساختار یک شبکه‌عصبی پیچش، لایه‌های ورودی، پیچش، تابع فعال‌ساز، انتخاب و تمام متصل وجود دارند. در این نوع از شبکه‌ها باید تعداد لایه‌های پیچش، لایه‌های انتخاب، اندازه فیلترها، تعداد نورون‌ها و نحوه اتصالات بین لایه‌ها مشخص شود. برای آموزش شبکه پیچش همانند شبکه‌های عصبی معمولی از روش انتشارخطا رو به عقب استفاده می‌شود [۸۷].

اولین کاربردهای موفقیت‌آمیز شبکه‌های عصبی کانولشن توسط آقای یان‌لی در سال ۱۹۹۰ توسعه داده شدند [۹۳]. از شبکه‌های معرفی شده معماری LeNet بهترین عملکرد را برای خواندن کدهای پستی و ارقام نشان داد [۹۳]. در سال‌های بعد، معماری AlexNet محبوبیت زیادی را در کاربردهای بینایی ماشین به دست آورد. این معماری توسط نویسندگان مرجع [۸۷] توسعه داده شد. این معماری در مسابقات شناسایی بصری مجموعه‌داده بزرگ ImageNet[۱] در سال ۲۰۱۲ ارائه شد و توانست بهترین عملکرد را در آن سال داشته باشد. معماری این شبکه، شبیه به معماری LeNet است ولی عمیق‌تر و بزرگ‌تر بوده و از ترکیب چندین لایه پیچش با هم استفاده می‌کند.

در معماری LeNet به ترتیب از لایه‌های پیچش و انتخاب، پیچش و انتخاب و سه لایه تمام متصل استفاده شد. در شبکه AlexNet به ترتیب از لایه‌های پیچش، انتخاب، دو پیچش، دو انتخاب و دو لایه‌ی تمام متصل استفاده شد. در واقع در شبکه‌های قدیمی‌تر، قرارگرفتن لایه انتخاب بعد از پیچش امری رایج بوده است. در روش پیشنهادی برای دسته‌بندی تصاویر نامتعارف از ترکیب دو معماری پایه AlexNet و LeNet استفاده شده که در ادامه معماری پیشنهادی تشریح می‌شود.

[۱] ImageNet Large Scale Visual Recognition Competition (ILSVRC)

۳ - ۲ - ۳معماری پیشنهادی برای شبکه‌های عصبی پیچش در تصاویر نامتعارف (*نوآوری*)

طراحی معماری شبکه در عملکرد سیستم پیشنهادی بسیار تاثیرگذار است. در روش پیشنهادی، معماری جدید با ترکیب معماری AlexNet و LeNet ارائه شد که در آن از لایه‌های پیچش، انتخاب و تمام متصل استفاده می‌شود. بخش‌هایی از انتخاب هر لایه مانند AlexNet است و از چند لایه پیچش پشت‌سرهم با تابع فعال‌ساز ReLU استفاده شده است. در این معماری، بخش‌هایی شبیه LeNet است و بعد از هر لایه‌ی پیچش یک لایه انتخاب در نظر گرفته شده است. به صورت تجربی این معماری عملکرد خوبی در شناسایی تصاویر نامتعارف دارد. در شبکه عصبی پیچش، در هر لایه پیچش ویژگی‌های بامعنا استخراج می‌شوند. در لایه‌های ابتدایی ویژگی‌های سطح پایین و از ترکیب این ویژگی‌ها در لایه بالاتر، ویژگی‌های معنایی سطح بالاتر که به مفاهیم موجود در تصویر نزدیک می‌باشد، استخراج می‌گردد. شکل ۳-۱ ساختار معماری پیشنهادی را نشان می‌دهد. معماری پیشنهادی، به صورت رابطه (۳-۱) می‌باشد. همان‌طور که در رابطه (۳-۱) نشان داده شده است، معماری پیشنهادی شامل چندین قسمت می‌باشد. دو قسمت اول لایه‌های پیچش، تابع فعال‌ساز ReLU و انتخاب را شامل می‌شود، قسمت بعدی دارای لایه پیچش به همراه تابع فعال‌ساز ReLU می‌باشد، قسمت بعدی لایه‌های پیچش، تابع فعال‌ساز ReLU و انتخاب را دربر می‌گیرد و سه قسمت آخر هم شامل لایه تمام متصل با تابع فعال‌ساز ReLU هستند.

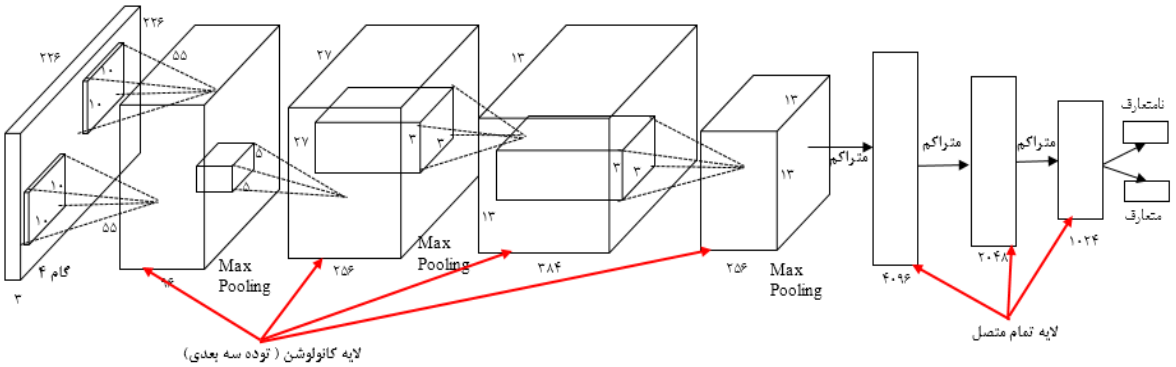
$$\begin{aligned} input \rightarrow [ConV \rightarrow ReLU \rightarrow Pooling] \times 2 \rightarrow [ConV \rightarrow ReLU] \\ \rightarrow [ConV \rightarrow ReLU \rightarrow Pooling] \rightarrow [FC \rightarrow ReLU] \times 3 \\ \rightarrow Output \end{aligned} \qquad (۱-۳)$$

در شبکه‌ی پیشنهادی، اندازه‌ی تصویر ورودی ۳×۲۲۶×۲۲۶ و اندازه‌ی فیلترهای استفاده شده در آن، ۳×۱۰×۱۰ به همراه گام[1] ۴ است. بلوک‌ها با اندازه ۳×۱۰×۱۰ پیکسل به یک بردار ستونی با اندازه ۳۰۰=۳×۱۰×۱۰ تبدیل می‌شوند. برای محاسبه اندازه (مکانی) توده خروجی از رابطه (۳-۲) استفاده می‌شود [۸۷].

$$\frac{W - F}{S} + 1 \qquad (۲-۳)$$

[1] Stride

در رابطه (۳-۲)، W اندازه تصویر ورودی، F اندازه ناحیه ادراکی نورون‌های لایه پیچش و S اندازه گام هستند. در این تحقیق، W، F و S به ترتیب ۲۲۶، ۱۰ و ۴ هستند. در صورتی که اندازه ناحیه ادراکی کوچک انتخاب شود ویژگی‌های بهتری را می‌توان استخراج نمود در همین راستا در معماری پیشنهادی اندازه این ناحیه ۱۰ در نظر گرفته شده که نسبت به معماری AlexNet کوچکتر می‌باشد. ۲۲۶ اندازه تصویر ورودی می‌باشد. با تکرار این عمل بر روی ورودی، سرانجام ماتریس خروجی با [۳۰۰×۳۰۲۵] تولید شد. هر ستون این ماتریس یک ناحیه ادراکی است و



شکل ۳-۱ نمای کلی از معماری پیشنهادی برای شبکه‌عصبی پیچش

تعداد ۵۵×۵۵=۳۰۲۵ ستون در کل وجود دارد. نواحی ادراکی با هم اشتراک دارند. از این رو، امکان تکرار چندین باره‌ی هم‌پوشانی هر عدد در توده خروجی وجود دارد.

به همین ترتیب، وزن‌های شبکه پیچش به بردارهای سطری تبدیل می‌شوند. در لایه بعدی، ۹۶ فیلتر (مجموعه وزن‌ها) با اندازه ۳×۱۰×۱۰ وجود دارد که تعداد وزن‌ها برابر با تعداد عناصر موجود در ناحیه ادراکی است. یعنی در ناحیه ادراکی با اندازه‌ی ۳×۱۰×۱۰، تعداد وزن‌ها نیز همین مقدار خواهد بود. این عمل باعث ایجاد ماتریس جدید با اندازه [۹۶×۳۰۰] خواهد شد.

نتیجه عملیات پیچش (ضرب نقطه به نقطه) برابر با اجرای یک ضرب ماتریسی بزرگ است. این عملیات نتیجه ضرب نقطه‌ای بین تمام فیلترها و تمام نقاط نواحی ادراکی را تولید می‌کند. خروجی این عملیات در شبکه، ماتریسی با اندازه [۹۶×۳۰۲۵] است که نتیجه ضرب نقطه‌ای بین هر فیلتر در هر موقعیت است. نتیجه نهایی به صورت ۹۶×۵۵×۵۵ تغییر شکل داده می‌شود. لایه تابع فعال‌ساز، بر روی تک‌تک

نورون‌ها، یک تابع فعال‌ساز را اعمال می‌کند. این تابع، اندازه‌ی توده مرحله قبل را تغییر نمی‌دهد. تابع فعال‌سازی ReLU با ورودی x، در رابطه (۳-۳) نشان داده شده است [۸۷].

$$f(x) = Max(0, x) \qquad (۳-۳)$$

استفاده از تابع فعال‌ساز ReLU، سرعت همگرایی گرادیان نزولی تصادفی را نسبت به تابع سیگموئید و تانژانت هیپربولیک بهبود داده و پیاده‌سازی این روش نسبت به تابع‌های دیگر ساده‌تر است. معمولاً بر روی خروجی تابع فعال‌ساز، نرمال‌سازی داده‌ها انجام می‌شود. اما تابع فعال‌ساز ReLU خواص مطلوبی نیز دارد و نیازی به نرمال کردن ورودی برای اجتناب از اشباع‌شدن[1] ندارد [۸۷]. با این وجود، برای تعمیم‌پذیری بیشتر این الگوریتم، از رابطه‌ی نرمال‌سازی (۳-۴) استفاده شده است [۸۷].

$$b_{x,y}^{i} = a_{x,y}^{i} / \left(k + \alpha \sum_{j=\max\left(0, i-\frac{n}{2}\right)}^{\min\left(N-1, i+\frac{n}{2}\right)} \left(a_{x,y}^{j}\right)^{2} \right)^{\beta} \qquad (۴-۳)$$

در رابطه‌ی (۳-۴)، $a_{x,y}^{i}$ فعالیت[2] نورون محاسبه شده توسط کرنل i در مکان (x,y) است و $b_{x,y}^{i}$ فعالیت نرمال‌شده است. N تعداد کرنل‌ها در این لایه و β، α، n، k مقادیر ابرپارامتر هستند که مانند مرجع [۸۷] از مجموعه اعتبارسنجی استخراج می‌شوند. در این تحقیق، $\beta = 0.75$، $\alpha = 10^{-4}$، $n = 5$ و $k = 2$ در نظر گرفته می‌شود.

رایج‌ترین شکل استفاده از لایه انتخاب با فیلترهایی با اندازه ۲×۲ به همراه گام ۲ است که هر برش عمقی در ورودی را با حذف ۲ عنصر از عرض و ۲ عنصر از ارتفاع کاهش می‌دهد و باعث حذف ۷۵٪ مقادیر موجود در آن برش عمقی می‌شود. لایه انتخاب تصویر ورودی را در هر برش عمقی به صورت مستقل (یعنی هر برش بدون توجه به برش دیگر) از لحاظ مکانی کاهش[3] می‌دهد. در بخش قبل ذکر شد که تابع‌های انتخاب زیادی برای شبکه‌های پیچش معرفی شد که معروف‌ترین آنها تابع انتخاب

[1] Saturating
[2] Activity
[3] Downsample

ماکزیمم است. این تابع یک تابع کاهش نمونه غیرخطی است و از هر چهار عنصر، ماکزیمم مقدار را حفظ می‌کند.

لایه‌های بعدی مانند یک شبکه‌عصبی پرسپترون دارای اتصالات تمام متصل هستند که هر لایه شامل ۴۰۹۶، ۲۰۴۸ و ۱۰۲۴ نورون می‌باشند. در نهایت لایه خروجی حاوی امتیاز دسته‌ها است که در این کاربرد لایه خروجی، امتیاز دو دسته متعارف و نامتعارف را شامل می‌شود.

در روش پیشنهادی برای دسته‌بندی تصاویر نامتعارف یک معماری نوین بر پایه معماری پایه AlexNet و LeNet ارائه شده است. این معماری، ویژگی‌های ساختاری دو معماری پایه را دارد که سبب بهبود عملکرد این معماری می‌شود. نتایج بیانگر این موضوع است که روش پیشنهادی عملکرد بهتری نسبت به روش‌های موجود در پایگاه داده استاندارد تصاویر نامتعارف دارد.

۳ - ۳ سیستم‌های دسته‌بندی تصویر بر اساس محتوا

در فصل قبل، انواع روش‌ها و رویکردهای ارائه شده در دسته‌بندی تصاویر و روش‌های استخراج اشیاء از تصویر مرور شد. مشکلات پیش رو در شناسایی و تشخیص نیز بررسی شد. اشاره شد که در این نوع از سیستم‌ها وجود اشیاء زیاد در تصویر، تغییرات شدت روشنایی، به دست آوردن رابطه بین اشیاء موجود در تصویر و تحلیل محتوا تصویر عمل دسته‌بندی تصویر را پیچیده و سخت کرده است. در همین راستا پیشنهاد می‌شود که با تحلیل ویژگی‌های سطح بالا و با به دست آوردن اشیاء موجود در تصویر به عنوان ورودی سیستم‌های تشخیص موضوع، عمل دسته‌بندی انجام شود. در گام تشخیص اشیاء، از روش یادگیری عمیق استفاده می‌شود. در گام بعدی با استخراج ویژگی‌های موضوعی نهفته در تصویر، عمل دسته‌بندی انجام می‌شود. توضیح مراحل مختلف کار در قسمت‌های بعدی ارائه می‌شود.

۳ - ۴ روش پیشنهادی برای دسته‌بندی صحنه‌های پیچیده

در این رساله دو روش پیشنهادی برای دسته‌بندی تصاویر با محیط‌های پیچیده معرفی می‌شود. در روش نخست از روش‌های سنتی یادگیری ماشین استفاده می‌شود. این روش عملکرد مناسبی از لحاظ دقت داشته ولی با مشکل پیچیدگی زمانی روبرو است و در مجموعه داده‌های با حجم بالا و تعداد کلاس زیاد عملکرد مناسبی از لحاظ زمان آموزش و آزمون از خود نشان نمی‌دهد. منظور از روش‌های سنتی، روش‌هایی می‌باشند که قبل از یادگیری عمیق محبوبیت و کاربرد زیادی داشتند ولی با معرفی شبکه‌های عمیق، کاربرد این روش‌ها در سال‌های اخیر کم رنگتر شده است. در این قسمت، روشی ترکیبی از Boosting و نواحی کاندیدا برای تشخیص اشیاء معرفی شد. در ادامه با توجه به برچسب نواحی و ویژگی‌های استخراج شده از نواحی انتخابی از SVM برای دسته‌بندی صحنه تصویر استفاده می‌شود.

در روش پیشنهادی دوم، یک سیستم جامع برای تشخیص و برچسب‌زنی اشیاء موجود در تصویر، استخراج مدل موضوعی نهفته در تصویر و شناسایی دسته صحنه تصویر ارائه می‌شود. این روش عملکرد مناسبی از لحاظ زمان اجرا و دقت از خود نشان می‌دهد. روش پیشنهادی بر پایه مفاهیم معنایی سطح بالا می‌باشد. در بخش بعدی این روش به تفصیل بیان خواهد شد.

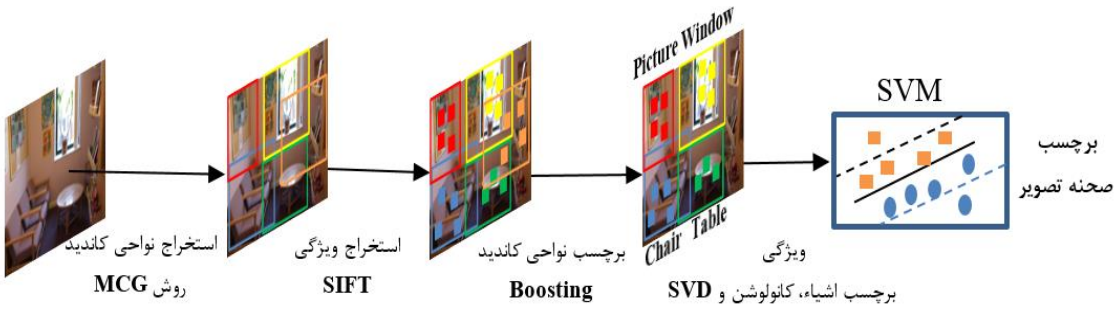
۳ - ۴ - ۱ روش پیشنهادی بر پایه Boosting و نواحی کاندیدا (*نوآوری*)

دسته‌بندی صحنه‌ی تصویر یکی از چالش‌های اساسی و مهم در بینایی ماشین و یادگیری ماشین می‌باشد. در این بخش، یک راهکار نوین برپایه ویژگی‌های سطح بالای استخراج شده از نواحی مرتبط با اشیاء به منظور دسته‌بندی صحنه تصویر ارائه می‌شود. در گام نخست نواحی کاندیدا شی توسط روش گروه بندی ترکیبی بر پایه چندین مقیاس[۱](MCG) شناسایی می‌شود. سپس برای استخراج ویژگی از روش پیچش و تجزیه مقادیر منفرد[۲](SVD) استفاده شده است. در گام بعدی از شبکه Boosting برای

[۱] Multiscale Combinatorial Grouping (MCG)

[۲] Singular Value Decomposition (SVD)

شناسایی برچسب نواحی و همچنین میزان شی بودن نواحی پیشنهادی استفاده شد. در نهایت با ویژگی‌های استخراج شده از نواحی مرتبط به شی و برچسب آن‌ها، دسته‌بندی تصویر توسط ماشین بردار پشتیبان انجام می‌شود. معماری پیشنهادی برای دسته‌بندی تصویر بر اساس روش‌های سنتی در شکل ۳-۲ نشان داده شده است.



شکل ۳-۲ شمای کلی از روش پیشنهادی بر پایه روش Boosting و ناحیه کاندیدا

۳ - ۴ - ۲روش Boosting

روش Boosting معرفی شده در مرجع [۹۴, ۹۵]، یک راه حل ساده برای تطبیق ترتیبی مدل‌های افزایشی مطابق رابطه (۳-۵) ایجاد می‌کند.

$$H\ (v,c) = \sum_{m=1}^{M} h_m(v,c) \qquad (۵-۳)$$

در اینجا c برچسب کلاس و v بردار ویژگی ورودی می‌باشد. h_m یادگیرنده ضعیف می‌باشد که در بسیاری از منابع به صورت رگرسیون به فرم $h_m = a\delta(v^f > \theta) + b$ بیان شده است که v^f نشان دهنده fامین جز (بعد) از بردار ویژگی v می‌باشد. θ نشان دهنده حد آستانه، δ تابع، a و b پارامترهای رگرسیون می‌باشند. Bossting اصلی طراحی شده برای انطباق در دسته‌بند دو کلاسه می‌باشد. دو روش رایج برای توسعه Boosting برای چند کلاسه وجود دارد. ساده‌ترین روش آموزش چندین دسته‌بند دوکلاسه به صورت جدا می‌باشد. یک رویکرد دیگر این است که اطمینان حاصل شود وزن در هر مرحله به صورت توزیع احتمالی نرمال در تمام طبقات تعریف می‌شود. در مرجع [۹۶] یک روش برای اشتراک یادگیرنده

ضعیف بر اساس کلاس خاص معرفی شد. برای مثال برای سه کلاس، کلاس‌بندها به فرم (۳-۶) تعریف می‌شود.

$$\begin{aligned} H\ (v,1) &= G^{1,2,3}(v) + G^{1,2}(v) + G^{1,3}(v) + G^{1}(v) \\ H\ (v,2) &= G^{1,2,3}(v) + G^{1,2}(v) + G^{2,3}(v) + G^{2}(v) \\ H\ (v,3) &= G^{1,2,3}(v) + G^{1,3}(v) + G^{2,3}(v) + G^{3}(v) \end{aligned} \qquad (۳-۶)$$

در رابطه (۳-۶) هر $G^{S(n)}(v)$ یک مدل افزودنی از فرم $G^{S(n)}(v) = \sum_{m=1}^{M_n} h_m^n(v)$ است. N به یک گره در گراف اشتراک اشاره می‌کند که مشخص می‌کند کدام تابع می‌تواند بین دسته‌بندها به اشتراک گذاشته شود. S(n) یک زیرمجموعه از دسته‌هایی می‌باشد که گره n را به اشتراک گذاشته است. ایده الگوریتم برای تشخیص اشیاء این است که در هر دور افزایش، زیرمجموعه‌های مختلف از کلاس‌ها بررسی می‌شود ($S \subseteq C$). با در نظر گرفتن انطباق یک کلاس‌بند ضعیف تمایز زیرمجموعه از پس‌زمینه در نظر گرفته می‌شود و زیرمجموعه‌ای انتخاب می‌گردد که حداکثر کاهش را در خطای مجموعه آموزش به ازای تمام کلاس‌ها به وجود می‌آورد. سپس، بهترین یادگیرنده ضعیف h(v,c) به یادگیرندهای قوی H(v,c) برای تمام کلاس‌های $c \in S$ اضافه می‌شود و توزیع وزن آنها به‌روز می‌شود. برای بهینه سازی تابع هزینه چند کلاسه از رابطه (۳-۷) استفاده می‌شود.

$$J = \sum_{c=1}^{C} E[e^{-z^c H(v,c)}] \qquad (۳-۷)$$

z^c برچسب عضویت (± 1) برای کلاس c می‌باشد. عبارت $z^c H(v,c)$ ، "حاشیه " نامیده می‌شود و وابسته به خطای تعمیم یافته می‌باشد. نسخه Boosting استفاده شده در این رساله بر اساس نسخه ارائه شده در مرجع [۹۵] با عنوان "gentleboost" است. دلیل انتخاب این روش سادگی در پیاده سازی، مقاومت عددی این روش می‌باشد همچنین به صورت تجربی در مرجع [۹۷] نشان داده شد که این روش در تشخیص چهره عملکرد خوبی دارد. بهینه سازی J توسط گام تطبیق نیوتن [۹۵] صورت می‌گیرد که این روش وابسته به حداقل مربع خطای وزن در هر گام است. در هر گام m، تابع H با رابطه (۳-۸) بروز رسانی می‌شود.

$$H(v,c) \coloneqq H(v,c) + h_m(v,c) \qquad (۸-۳)$$

در این رابطه، h_m طوری انتخاب می‌شود (رابطه (۳-۹) که تقریب تیلور درجه دوم از تابع هزینه حداقل شود:

$$\underset{h_m}{argmin} J(H + h_m) \cong \underset{h_m}{\operatorname{argmin}} \sum_{c=1}^{C} E\left[e^{-z^c H(v,c)}(z^c - h_m)^2\right] \qquad (۹-۳)$$

برای هر نمونه i و کلاس c، وزن‌ها به صورت $w_i^c = e^{-z^c H(v_i,c)}$ تعریف شد. این کاهش برای حداقل کردن خطای مربع وزن به صورت رابطه (۳-۱۰) محاسبه شد. تابع بهینه مطلوب h_m در مرحله m به صورت رابطه (۳-۱۱) تعریف شد.

$$J_{wse} = \sum_{c=1}^{C} \sum_{i=1}^{N} w_i^c \left(z_i^c - h_m(v_i,c)\right)^2 \qquad (۱۰-۳)$$

$$h_m(v,c) = \begin{cases} a\delta\left(v_i^f > \theta\right) + b & if \ \in S(n) \\ k^c & if \ c \notin S(n) \end{cases} \qquad (۱۱-۳)$$

۳ - ۴ - ۳ روش MCG برای ناحیه کاندیدا

همانطور که در فصل قبل اشاره شد، سیستم تشخیص اشیاء به الگوریتم‌های ناحیه‌های کاندیدا برای تخمین مکان اشیاء بسیار وابسته است. در این روش‌ها نمی‌توان سراسر تصویر را به دلیل محدودیت‌های زمانی و حالت‌های مختلف از اشیاء جستجو کرد. در همین راستا، ابتدا باید ناحیه‌های کاندیدا وجود شیء استخراج شوند. در گام بعدی با توجه به این نواحی تشخیص اشیاء انجام گیرد.

با توجه به نتایج مقایسات انجام شده در این رساله (فصل قبل)، روش گروه‌بندی ترکیبی بر پایه چندین مقیاس یک الگوریتم سریع برای محاسبه قطعه‌بندی سلسله مراتبی چندین مقیاسه است که بر روی مجموعه داده PASCAL VOC 2012 و BSDS500 عملکرد مناسبی در تشخیص اشیاء از خود نشان داده است. در همین راستا در روش پیشنهادی از این روش برای استخراج نواحی کاندیدا استفاده می‌شود.

۳ - ۴ - ۴ استفاده از Boosting در تشخیص اشیاء

در گام نخست از روش پیشنهاد ناحیه کاندیدا MCG برای نواحی که می‌توانند به عنوان شی معرفی شوند استفاده شد. در هر گام تعداد زیادی وصله[۱] به صورت تصادفی از نواحی کاندیدا برای اشیاء موجود استخراج شدند. سپس واژه‌نامه‌ای از تمام وصله‌های تصویر ساخته شد. در این روش، اندازه اشیاء نرمال و در پنجره مربعی ۳۲*۳۲ پیکسل در نظر گرفته شد. برای هر وصله، یک ماسک خاص که مشخص کننده مکانی که patch از تصویر اصلی استخراج شده در نظر گرفته شد. در این رساله، به ازای هر وصله تولید شده، بردار ویژگی برای هر تصویر محاسبه شد. بردار ویژگی محاسبه شده در مکان x و مقیاس σ توسط رابطه (۳-۱۲) محاسبه شد.

$$v^f(x,\sigma) = \left(w_f * \left|I_\sigma \otimes g_f\right|^p\right)^{\frac{1}{p}} \qquad (۱۲-۳)$$

در این رابطه $\otimes$ نمایش‌دهنده همبستگی نرمال شده بین نواحی کاندید تصویر I_σ و فیلتر g_f است و * عملگر پیچش است. $w_{f(x)}$ پنجره/ماسک فضایی است. $v(x,\sigma)$ بردار ویژگی محاسبه شده در مکان xو مقیاس σ است و v^f نشان دهنده fامین مولفه بردار است. ویژگی‌های مربوط به مقیاس‌های مختلف σ توسط تغییر مقیاس تصویر محاسبه می‌شوند. توان p سبب تولید ویژگی‌ها با نوع‌های مختلف می‌شود. برای نمونه p=1 بردار میانگین را توسط فیلتر محاسبه می‌کند که این ویژگی برای محاسبه بافت مناسب است. برای p > 10 بردار ویژگی به $v^f \cong \max_{x\epsilon S_w}\{\left|I_\sigma \otimes g_f\right|\}$ تبدیل می‌شود که $S_{w(x)}$ پشتیبان پنجره برای یک ویژگی در مکان x است که برای انطباق الگو بسیار مناسب می‌باشد. با تغییر ماسک فضایی، می‌توان اندازه و مکان ناحیه که ویژگی در آن ارزیابی شده را تغییر داد. با این کار می‌توان ویژگی‌هایی تولید کرد که به خوبی مکان‌یابی شده‌اند. در زمان آموزش، هزاران وصله از تصاویر آموزش استخراج شد که پیچش از هر وصله محاسبه شد. هزینه محاسبات برای هر وصله، توسط رابطه (۳-۱۳) با محاسبات خطی از فیلتر یک بعدی جداگانه کاهش یافت.

[۱] Patch

$$\hat{g}_f = \sum_{n=1}^{r} u_n v_n^T \qquad (۱۳-۳)$$

در رابطه (۳-۱۳)، u_n و v_n یک فیلتر یک بعدی هستند و r نشان دهنده میزان تقریب است. این تجزیه با اعمال SVD به ماتریس g_f محاسبه شد. برای هر فیلتر، مقدار r به صورتی انتخاب شد که $|\hat{g}_f^T\ g|$ بزرگتر از ۰/۹۵ باشد.

۳ - ۴ - ۵ دسته بندی صحنه تصویر با روش پیشنهادی برپایه Boosting

برای دسته‌بندی صحنه از ویژگی‌های استخراج شده از نواحی مرتبط به اشیاء که توسط Boosting به عنوان شی معرفی شد برای دسته‌بندی صحنه استفاده شد. برای دسته‌بندی صحنه از روش ماشین بردار پشتیبان استفاده شد. در واقع در روش پیشنهادی از نواحی مرتبط با شی برای استخراج ویژگی و دسته‌بندی استفاده می‌کند و نواحی مرتبط با پس‌زمینه در نظر گرفته نشده است.

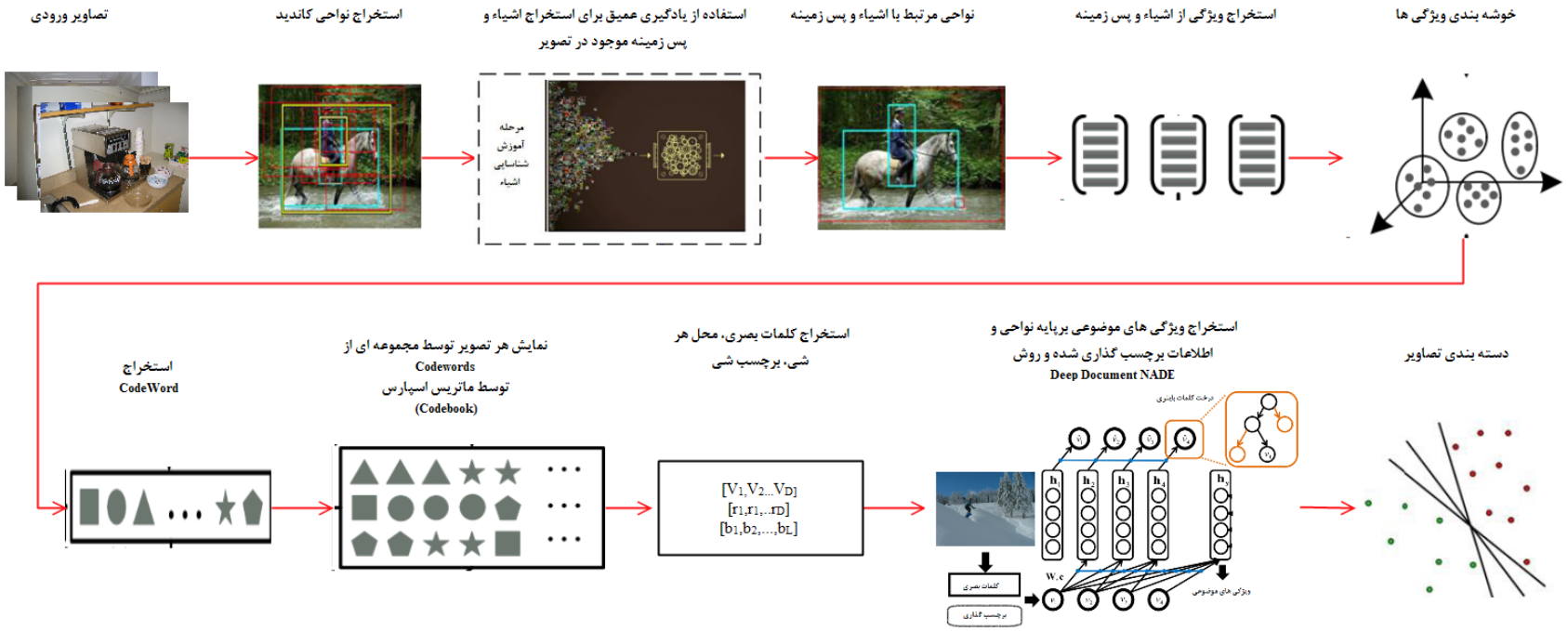
روش ارائه شده در این بخش از لحاظ دقت روش مناسبی است ولی زمان اجرای این روش بسیار بالا است. در همین راستا در ادامه روشی نوین بر پایه شبکه عصبی عمیق و مدل‌های موضوعی معرفی می‌شود. این روش علاوه بر اینکه عملکرد مناسبی از لحاظ دقت دارد زمان اجرای آن نیز مناسب می‌باشد.

۳ - ۵ روش پیشنهادی بر پایه مدل موضوعی

در مسئله طبقه‌بندی تصویر، تعدادی از تصاویر رنگی برای آموزش طبقه‌بند و تعدادی دیگر از تصاویر برای آزمون طبقه‌بند استفاده می‌شوند. ابتدا یک طبقه‌بندی اولیه روی کل مجموعه آموزش داده و هر تصویر متعلق به یکی از دسته‌ها اعمال می‌شود. سپس هر دسته به عنوان برچسب داده معرفی می‌شود. سیستم طبقه‌بند با آموزش روی داده‌های آموزشی، برچسب مربوط به هر منطقه از داده‌ها را مشخص می‌کند. در ادامه هر تصویر آزمون در یکی از مناطق مربوط به داده‌ها قرار گرفته و طبقه‌بند برچسب آن را مشخص می‌کند.

همانطور که در بخش ۱-۲ هدف رساله بیان شد $X = \{x_1, x_2, \dots, x_N\}$ تصاویر مجموعه آموزش و $Y = \{y_1, y_2, \dots, y_N\}$ برچسب‌های این تصاویر می‌باشد. هدف سیستم تشخیص برچسب داده‌های آزمون است.

چارت روش پیشنهادی در این رساله در شکل ۳-۳ نشان داده شده است که در ادامه این روش تشریح می‌شود.



شکل ۳-۳ روش پیشنهادی رساله برای دسته بندی تصاویر

در روش پیشنهادی مراحل زیر روی هر تصویر ورودی x_i انجام می‌شود. این تصاویر متعلق به مجموعه آموزش هستند:

۱. در گام نخست با استفاده از روش نواحی کاندیدا، نواحی کاندید اشیاء موجود در تصویر، استخراج می‌شوندکه دراینجا این نواحی به عنوان $PR_{i.k} = \{rp_{i,1}, rp_{i,2}, \dots, rp_{i,m}\}$ معرفی می‌شوند. در بسیاری از کارهای انجام شده در سال‌های اخیر تعداد نواحی کاندیدا ثابت فرض می‌شود ولی در روش پیشنهادی این محدودیت وجود ندارد. در گام‌های بعدی این تعداد کاهش پیدا می‌کند.

۲. در گام بعدی یک شبکه عمیق برای شناسایی اشیاء موجود در تصویر آموزش داده می‌شوند. در این قسمت از مجموعه ImageNet برای وزن‌دهی اولیه یادگیری عمیق استفاده می‌شود [۸۷]. این مجموعه داده به عنوان یک مجموعه داده بزرگ با تعداد اشیاء زیاد می‌باشد. در مرحله آزمون، اشیاء موجود در تصویر i با توجه به نواحی کاندیدا استخراج می‌شوند. این نواحی کاندیدا با عنوان RP در نظر گرفته شده است.

در روش پیشنهادی از یک شبکه عصبی عمیق یکپارچه برای انجام دو مرحله قبل استفاده می‌شود. برای کاهش نواحی کاندیدا پیشنهادی از یک شبکه ترکیبی برپایه SVM فازی دو کلاسه برای جدا سازی اشیاء از پس‌زمینه و SVR برای برچسب‌زنی به نواحی استفاده شده است. جزئیات روش پیشنهادی در ادامه به تفصیل بیان می‌شود.

۳. در مرحله بعدی، نواحی مربوط به RP با در نظر گرفتن میزان شی بودن آن انتخاب می‌شوند. تعداد نواحی انتخاب شده کمتر از m است. در کارهای انجام شده در سال‌های اخیر، پس‌زمینه به عنوان یک شی در نظر گرفته می‌شود ولی در روش پیشنهادی توسط یک فاز مجزا پس‌زمینه از شی جدا می‌شود. این عمل سبب کاهش نواحی کاندیدا می‌شود. در واقع خروجی حاصل از این مرحله، به نواحی اشیاء با عنوان $PO_{i.t} = \{rc_{i,1}, rc_{i,2}, \dots, rc_{i,q}\}$ تبدیل می‌شود. در مراحل بعدی نواحی RO به عنوان نواحی کل تصویر در نظر گرفته می‌شود.

در کارهای انجام شده در سال‌های اخیر، کل تصویر ورودی به عنوان ورودی سیستم‌های استخراج مدل‌های موضوعی استفاده شده است، ولی در اینجا از نواحی مربوط به اشیاء استفاده می‌شود و نواحی معنایی بهتری برای استخراج موضوع استفاده می‌شود. از طرفی ویژگی‌هایی مانند SIFT، تعدادی ویژگی را در کل تصویر در نظر می‌گیرد که این ویژگی‌ها می‌توانند از نواحی با ارزش کمتر باشد و ویژگی‌های بامعنای مربوط به شی را از دست دهند. اما در روش پیشنهادی، ویژگی مربوط به ناحیه‌های مرتبط به شی استخراج می‌شوند. در روش پیشنهادی، نواحی مربوط به پس‌زمینه جدا شده است و فقط نواحی مرتبط به شی استفاده می‌شود. دلیل این امر، مشابه بودن پس‌زمینه‌ها در کلاس‌های مشابه می‌باشد. برای نمونه در تصویری با شی غالب "اسب" و پس‌زمینه "چمن" (صحنه اسب سواری) دسته‌ای متفاوت از تصویر با شی غالب "انسان" با پس‌زمینه "چمن" (صحنه بازی گلف) دارد. در این روش از ترکیب نوینی از SVM فازی دو کلاسه و SVR برای ارزش‌دهی به نواحی استخراج شده استفاده می‌شود.

۴. در مرحله بعدی، از هر ناحیه تعدادی نقطه به همراه همسایگی مشخصی از آن انتخاب می‌شود که به آنها نقاط محلی می‌گویند. سپس تمامی این نواحی محلی با روش یکسان‌سازی هیستوگرام از نظر شدت نور و اندازه، یکسان‌سازی می‌شوند. با این روش، نواحی محلی نسبت به تغییرات نور و اندازه تا حد مناسبی مقاوم می‌شوند. سپس از نقاط محلی، ویژگی SIFT استخراج می‌شود. ویژگی SIFT ، هیستوگرامی از اندازه و جهت بردارهای گرادیان تصویر در نقاط تصویر است. این ویژگی نسبت به دوران تصویر مقاوم است. خروجی این الگوریتم یک بردار ۱۲۸ بعدی است که نسبت به تغییرات شدت نور، انداز و دوران تقریبا مقاوم است. در این تحقیق، یک بردار ویژگی SIFT به صورت $s_{i,h}$ نشان داده می‌شود که به معنی ویژگی SIFT استخراج شده از h-امین نقطه محلی تصویر iام است.

۵. در گام بعدی، بر روی بردارهای SIFT حاصل از تمامی تصاویر آموزش، الگوریتم خوشه‌بندی اجرا می‌شود و مرکز هر خوشه به عنوان ویژگی سرآمد انتخاب می‌شود. در واقع مرکز هر خوشه

به بهترین شکل، داده‌های موجود در یک خوشه را توصیف می‌کند. منظور از توصیف در روش‌های مختلف، متفاوت است و می‌تواند به معنی کم‌ترین خطای بازسازی یا نزدیکی فاصله‌ی اقلیدسی به داده‌ها باشد. در روش پیشنهادی از یک خوشه‌بندی توسعه یافته شده از k-means استفاده شده که نسبت به روش اصلی سریعتر و دقیق‌تر می‌باشد. lامین مرکز خوشه به صورت $cw_{i,l}$ نشان داده می‌شود. به هر مرکز خوشه، پایه و به مجموعه پایه‌ها واژه‌نامه گفته می‌شود. در واقع در این بخش کلمات کد استخراج می‌شود.

۶. در مرحله بعدی از روش کدگذاری استفاده می‌شود. به هر ویژگی $s_{i,h}$ یک کد $a_{i,h}$ اختصاص داده می‌شود. ابعاد بردار کد $a_{i,h}$ برابر با تعداد مراکز خوشه‌ها است و همه عناصر آن به جز یک عنصر برابر با صفر است. اگر فرض کنیم که داده $s_{i,h}$ به l امین خوشه تعلق داشته باشد عنصر l ام کد $a_{i,h}$ برابر با یک و باقی عناصر برابر با صفر هستند. با کمی تأمل مشخص می‌شود که در واقع بردار کد، یک نمایش سطح بالا از ویژگی‌های سطح پایین SIFT هستند. زیرا به جای استفاده از ویژگی سطح پایین، از ارتباط آن با ویژگی‌های برای نمایش آن استفاده می‌شود. در واقع در این مرحله کتابچه کد استخراج می‌شود و در ادامه بهترین نماینده از هر خوشه استخراج می‌شود.

برای محاسبه نواحی محلی از نواحی مرتبط به اشیاء استفاده شده است. در روش‌های ارائه شده از k-means توسعه داده شده برای بهبود سرعت روش مجموعه کلمات استفاده شده است. علاوه بر این از روش kd-tree برای محاسبه فاصله بین ویژگی‌های استخراج از کلمات بصری استفاده می‌گردد. در روش‌های موجود برای خوشه‌بندی، تعداد خوشه باید از قبل موجود باشد. ولی در روش پیشنهادی، می‌توان تعداد خوشه‌ها را با توجه به تعداد اشیاء موجود در نظر گرفت. همچنین کتابچه‌کد با توجه به نواحی جداگانه به دست می‌آید.

۷. در گام بعدی، کلمه‌های بصری استخراج می‌شوند. هر تصویر i می‌تواند توسط بسته از کلمات بصری $V_{i,f} = \{v_{i,1}, v_{i,2}, \dots, v_{i,D}\}$ توصیف شود. در این رابطه، هر $v_{i,f}$ ایندکس نزدکترین

خوشه k-means در f امین توصیفگر استخراج شده از تصویر i است و D تعداد توصیفگرهای استخراج شده است. همانطور که قبلا اشاره شد، اطلاعات مکانی در درک تصویر نقش مهمی دارند. در این قسمت مانند روش مشابه، حضور کلمات بصری و ناحیه‌ای که در تصویر ظاهر شده مدل می‌شوند. در این تحقیق، فرض شده که تصویر به چندین ناحیه جدا $R = \{R_1, R_2, \dots, R_M\}$ تقسیم می‌شود که M تعداد نواحی است. تصویر به صورت $V_{i,f}^R = [v_{i,1}^R, v_{i,2}^R, \dots, v_{i,D}^R] = [(v_{i,1}, r_{i,1}), (v_{i,2}, r_{i,2}), \dots, (v_{i,D}, r_{i,D})]$ نمایش داده می‌شود که $r_{i,f} \in R$ نشان دهنده نواحی متشکل از کلمه بصری $V_{i,f}$ است. در این مرحله، موقعیت مکانی و برچسب اشیاء به عنوان ورودی سیستم در نظر گرفته می‌شوند. اگر A واژگان همه‌ی کلمات برچسب‌زده شده باشد. برچسب‌گذاری بدست آمده از یک تصویر، به صورت $b = \{b_1, b_2, \dots, b_L\} \; and \; b \in A$ است و L نیز تعداد کلمات موجود در برچسب گذاری است. در واقع تصویر برچسب‌گذاری شده را می‌توان ترکیبی از مجموعه کلمات بصری و کلمات برچسب-گذاری شده دانست و به صورت $v_i^A = [v_{i,1}^A, \dots, v_{i,D}^A, v_{i,D+1}^A, \dots, v_{i,D+L}^A] = [v_{i,1}^R, \dots, v_{i,D}^R, b_{i,1}, \dots, a_{i,L}]$ نوشت.

در این بخش از برچسب استخراج شده از مرحله قبل توسط SVR برای کلمات بصری استفاده می‌شود. که این برچسب‌ها هم به عنوان ورودی سیستم استخراج مدل موضوعی و هم برای سیستم دسته‌بندی نهایی در نظر گرفته می‌شود. از نواحی استخراجی از SVM فازی دوکلاسه برای استخراج کلمات کلیدی استفاده می‌شود.

۸. در این مرحله، از ویژگی‌های استخراج شده برای استخراج ویژگی‌های موضوعی استفاده می‌شود. در این گام ویژگی‌های موضوعی مرتبط با هر تصویر استخراج می‌شوند. روش‌های بسیار زیادی برای استخراج ویژگی‌های موضوعی از متن معرفی شده‌اند. در سال‌های اخیر، این روش‌ها در کاربرد پردازش تصویر نیز به کار برده شده‌اند. یکی از این روش‌های قدرتمند، روش ارائه شده

در مرجع [۱] است که بر مبنای این روش، به ازای هر تصویر i به تعداد S ویژگی موضوعی استخراج و بردار ویژگی $Z_{i,s}$ تشکیل می‌شود.

در روش‌های موجود، برچسب به عنوان یک ورودی در نظر گرفته شده و باید از قبل مشخص باشد. ولی در روش پیشنهادی، اشیاء موجود توسط یادگیری عمیق استخراج می‌شوند و به عنوان ورودی سیستم مدل‌سازی موضوعی در نظر گرفته می‌شوند. در واقع از برچسب‌های استخراجی برای کلمات بصری استفاده می‌شود. با توجه به این که موقعیت نواحی استخراجی مرتبط به شی در تصویر مشخص می‌باشد می‌توان تصویر را به نواحی دقیقتر تقسیم نمود.

۹. در نهایت با استفاده از این ویژگی‌های استخراجی دسته-بندی صحنه تصویر انجام می‌شود.

در این مرحله، از ویژگی‌هایی مانند: ویژگی‌های موضوعی، نواحی و برچسب مرتبط به اشیاء (برای کلمات بصری) و ویژگی پیچش مورد استفاده قرار می‌گیرد. از طرفی می‌توان از ویژگی‌های هر شیء نیز بهره برد. در قسمت بعد این مراحل با جزئیات بیشتر توضیح داده می‌شوند.

۳ - ۵ - ۱ معماری روش پیشنهادی بر پایه مدل موضوعی

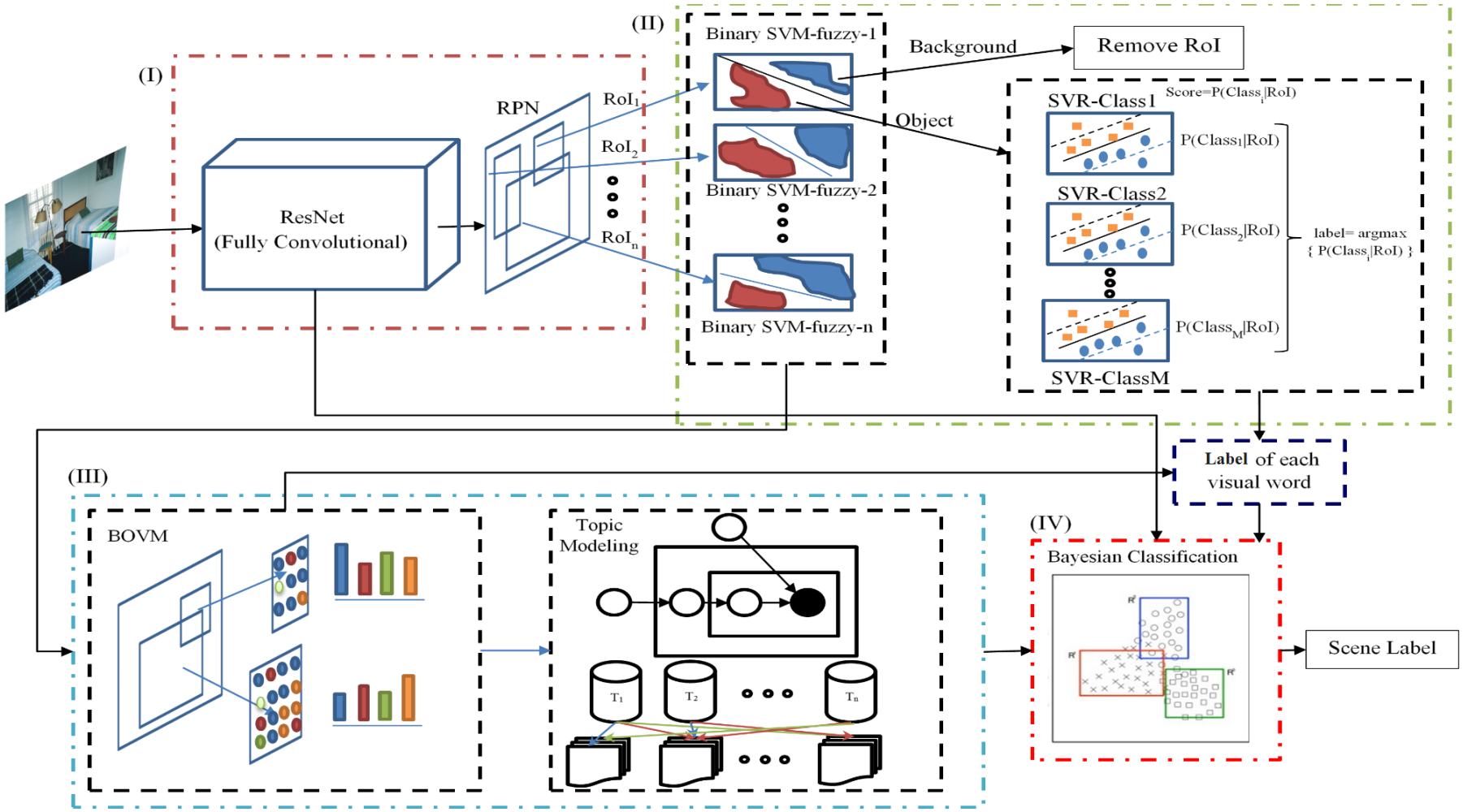
در این بخش جزئیات بیشتری از روش پیشنهادی بیان می‌شوند. در شکل ۳-۴ روش پیشنهادی نشان داده شده است.

همانطور که درشکل ۳-۴ مشخص است، روش پیشنهادی شامل چندین بخش می‌باشد. در حالت کلی می‌توان سه بخش اصلی برای روش پیشنهادی در نظر گرفت، بخش اول شامل انتخاب نواحی کاندیدا و مشخص نمودن اشیاء موجود در تصویر می‌باشد. بخش دوم، بدست آوردن موضوعات نهفته در تصویر بر اساس نواحی استخراج شده از مرحله قبل را دربر می‌گیرد. در نهایت دسته‌بندی تصویر بر اساس ویژگی‌های استخراج شده از کل تصویر بر پایه یادگیری عمیق، برچسب شی برای کلمات بصری و موضوعات نهفته در نواحی کاندیدا انجام می‌شود. هر قسمت در بخش‌های بعدی این فصل با جزئیات بیشتر بیان می‌شود.

در روش پیشنهادی تشخیص برچسب اشیاء شامل چندین فاز می‌باشد. در گام نخست نواحی کاندیدا از تصویر ورودی استخراج می‌شود. در این مرحله از معماری نوین بر اساس R-FCN استفاده شده است. در این روش از یک معماری تمام پیچش با عنوان ResNet برای استخراج نواحی کاندیدا پیشنهاد شده است. لازم به ذکر است که در معماری تمام پیچش از یک تابع زیان جدید استفاده شده که تاکنون برای شناسایی شی استفاده نشده است.

در این روش، بسیاری از نواحی کاندیدای انتخابی مناسب نمی‌باشند. منظور از مناسب نبودن، همپوشانی بیش از اندازه شی با پس‌زمینه و هم‌پوشانی نواحی چند شی با هم می‌باشد. علاوه بر این استخراج تعداد زیاد نواحی می‌تواند سرعت و دقت روش پیشنهادی را در مراحل بعدی کاهش دهد. در همین راستا، در این رساله یک معماری نوین بر اساس SVM فازی دو کلاسه و SVR پیشنهاد شده است. در این معماری، در گام نخست تمام نواحی کاندیدا به صورت خط لوله به عنوان ورودی SVM فازی دو کلاسه در نظر گرفته می‌شوند. در این مرحله خروجی شامل دو کلاس می‌باشد که تعیین می‌کند ناحیه ورودی شی یا پس‌زمینه است. لازم به ذکر است که SVM فازی دو کلاسه برای این دو کلاس آموزش داده می‌شود. در مرحله بعد نواحی که توسط این روش به عنوان شی انتخاب شدند، به عنوان ورودی SVR در نظر گرفته می‌شوند. به تعداد کلاس‌های مرتبط به اشیاء، SVR در نظر گرفته می‌شود که با استفاده از SVRها برچسب هر ناحیه تعیین می‌گردد. جزئیات این روش در بخش‌های بعدی توضیح داده می‌شود. در بخش دوم روش پیشنهادی، از نواحی خروجی SVM فازی دو کلاسه برای استخراج کلمات بصری و در ادامه استخراج موضوعات نهفته در تصویر استفاده می‌شود. جزئیات این روش در بخش‌های بعدی به تفصیل بیان می‌شود. در ادامه از یک روش مطرح برای استخراج موضوعات نهفته استفاده می‌شود. این روش در مرجع [۱] معرفی شده است. یک دلیل اساسی برای انتخاب این روش سرعت بالای این روش می‌باشد. برای استخراج ویژگی‌های موضوعی از لایه ماقبل آخر این روش استفاده می‌شود.

در بخش سوم روش پیشنهادی، از ویژگی‌های استخراج شده توسط لایه آخز شبکه عصبی تمام پیچش، برچسب شی (برای کلمات بصری) و موضوعات نهفته از تصویر برای دسته‌بندی استفاده می‌شود. برای دسته‌بندی از طبقه‌بند بیزین استفاده شده است.



شکل ۳-۴ روش پیشنهادی با بخش های مجزا

۳ - ۵ - ۲ استفاده از R-FCN با معماری جدید (نوآوری)

.همانطور که در فصل قبل توضیح داده شد، شبکه R-FCN یکی از جدیدترین شبکه‌های عمیق ارائه شده برای تشخیص اشیاء است. این شبکه از ترکیب شبکه RPN و شبکه تمام پیچش ساخته شده است. خروجی‌های این شبکه با توجه به موقعیت قرار گرفتن اشیاء، برای هر RoI امتیازدهی شده‌اند. در این روش از دو تابع زیان، آنتروپی متقابل و تخمین محدوده برای مکان و برچسب اشیا استفاده شده است. تابع زیان، آنتروپی متقابل برای طبقه‌بندی استفاده شده است. این تابع در رابطه (۳-۱۴) نشان داده شده است. تابع زیان استفاده شده در مرجع [۵۲] برای هر RoI، جمع شده و زیان تخمین محدوده به صورت رابطه (۳-۱۵) محاسبه می‌شود.

$$CEL = -\sum_j y^{*(j)} \log \sigma(O)^{(j)} \qquad (۱۴-۳)$$

$$L(S, t_{x,y,w,h}) = L_{cls}(S_{c^*}) + \lambda[c^* > 0] L_{reg}(t, t^*) \qquad (۱۵-۳)$$

در رابطه (۳-۱۴) y^* خروجی مطلوب، O خروجی شبکه و (j) نشان دهنده j-امین بعد بردار است. σ تخمین احتمالات را نشان می‌دهد. در رابطه (۳-۱۵)، c^* برچسب مطلوب RoI است. $c^* = 0$ به معنی پس‌زمینه است. در واقع $[c^* > 0]$ شرط مشخص کننده پس‌زمینه است. در بخش‌های غیر از پس‌زمینه، برابر با یک و در بقیه‌ی نواحی صفر است. هر RoI با همپوشانی بیش از ۵۰٪ با نواحی مطلوب، به عنوان نمونه مثبت در نظر گرفته می‌شود. t و t^* به ترتیب مقادیر مرتبط با ناحیه تخمینی و ناحیه مطلوب هستند. x, y, w, h نشان‌دهنده مختصات محدوده شی می‌باشند. $L_{cls}(S_{c^*}) = -\log(S_{c^*})$ زیان آنتروپی متقابل برای طبقه‌بندی است. تابع زیان تخمین محدوده بر اساس مرجع [۱۶] است. $L_{reg}(t, t^*)$ تابع زیان تخمین محدوده در رابطه (۳-۱۶) و (۳-۱۷) بیان شده‌اند.

$$L_{reg}(t, t^*) = \sum_{i \epsilon \{x,y,w,h\}} smooth_{L_1(t_i - t_i^*)} \qquad (۱۶-۳)$$

$$smooth_{L_{1(x)}} = \begin{cases} 0.5x^2 & if\ x < 1 \\ |x| - 0.5 & otherwise \end{cases} \qquad (۱۷-۳)$$

در این رابطه‌ها از تابع زیان $L_{1(x)}$ ستفاده شده که تابع زیان نرم یک است. دلیل استفاده از نرم یک حساسیت کمتر آن به داده‌های پرت در مقایسه با نرم دو (L_2) است. نرم دو در مرجع [۱۸] استفاده شده است.

استفاده از تابع زیان مناسب، در دقت شبکه بسیار تاثیرگذار است [۹۱]. به همین دلیل در این رساله برای بهبود شبکه R-FCN از یک تابع زیان جدید استفاده می‌شود. مرجع [۹۱]، دوازده تابع زیان معروف را معرفی و تحلیل کرده که این توابع قابلیت استفاده در شبکه‌های عمیق را دارند. نتایج حاصل نشان می‌دهد که استفاده از تابع زیان وابستگی زیادی به کاربرد مورد نظر دارد. در این رساله به تحلیل و مقایسه بین تابع زیان اختلاف کوشی-شوارتز و زیان آنتروپی متقابل پرداخته شده است. شایان ذکر است، این تابع زیان در شبکه‌های عمیق برای شناسایی اشیاء و روش R-FCN استفاده نشده است. آزمایش‌های انجام شده در مرجع [۹۱] نشان می‌دهند که تابع زیان اختلاف کوشی-شوارتز نسبت به زیان آنتروپی متقابل از لحاظ سرعت و عملکرد بهینه‌تر است. در همین راستا، به جای تابع زیان شبکه R-FCN اصلی (تابع زیان آنتروپی) از تابع اختلاف کوشی-شوارتز استفاده شده است. این تابع زیان به صورت رابطه (۳-۱۸) تعریف می‌شود [۹۸].

$$DCS = -\log\frac{\sum_j \sigma(o)^j\, y^{*(j)}}{\|\sigma(o)\|_2 \|y\|_2} \qquad (۱۸-۳)$$

در رابطه (۳-۱۸) y^* خروجی مطلوب، O خروجی شبکه، σ تخمین احتمالی و (j) نشان دهنده jامین بعد بردار است.

استفاده از این تابع زیان می‌تواند در بالا بردن دقت شبکه تمام پیچش موثر باشد. نتایج به دست آمده در این رساله نیز نشان می‌دهد که این تابع زیان برای تشخیص اشیاء موثر می‌باشد.

در روش پیشنهادی برای استخراج اشیاء از یک ماشین بردار پشتیبان فازی دو کلاسه استفاده می‌شود که در ادامه این روش توضیح داده می‌شود.

۳ - ۵ - ۳ ماشین بردار پشتیبان فازی دو کلاسه

ماشین بردار پشتیبان، به عنوان یک ابزار قدرتمند برای طبقه‌بندی و رگرسیون، در بسیاری از کاربردهای عملی مورد استفاده قرار گرفته شد [۹۹]. نسخه‌های بسیاری از ماشین بردار پشتیبان در سال‌های اخیر معرفی شده است که از معروف‌ترین آنها می‌توان به نسخه فازی اشاره نمود. اغلب SVM فازی برای حل مسائلی استفاده می‌شود که الگوهای متعلق به یک کلاس اغلب نقش‌های مهمتری در طبقه‌بندی دارند. در روش پیشنهادی بعد از مشخص نمودن نواحی مرتبط به اشیاء، میزان تعلق هر ناحیه به شی خاص مشخص می‌شود. در واقع در این گام از SVM فازی دوکلاسه برای برچسب‌زنی شی استفاده می‌شود. در این راستا از روش معرفی شده در مرجع [۱۰۰] برای طبقه‌بندی استفاده می‌شود. خروجی این طبقه‌بند دوکلاسه فازی مشخص کننده شی یا پس‌زمینه بودن است. در صورتی که پس‌زمینه تشخیص داده شود این ناحیه کنار گذاشته می‌شود ولی در صورت تشخیص وجود شی، در مرحله بعدی برچسب کلاس مشخص می‌شود. این عمل سبب کاهش عمل برچسب‌زنی تصویر می‌شود.

مجموعه آموزش برای مسئله طبقه‌بندی دوکلاسه به صورت رابطه (۳-۱۹) تعریف می‌شود:

$$T^* = \{(x_1, m_1), (x_2, m_2), \dots, (x_l, m_l)\} \qquad (۱۹-۳)$$

در رابطه (۳-۱۹)، x_i نمونه‌ها، $m_i \epsilon [0,1]$ عضویت فازی می‌باشد که میزان تعلق iامین مشاهده x_i به کلاس مثبت را ارزیابی می‌کند و l تعداد نمونه‌های مجموعه آموزش می‌باشد. در این روش p نمونه $\{(\widetilde{x_1}, \widetilde{m_1}), (\widetilde{x_2}, \widetilde{m_2}), \dots, (\widetilde{x_p}, \widetilde{m_p})\}$ از مشاهدات به عنوان مشاهدات مثبت و q نمونه $\{(\widehat{x_1}, \widehat{m_1}), (\widehat{x_2}, \widehat{m_2}), \dots, (\widehat{x_q}, \widehat{m_q})\}$ از مشاهدات به عنوان نمونه‌های منفی در نظر گرفته می‌شود. برچسب نمونه‌ها از رابطه $Y_i = 2m_i - 1$ محاسبه می‌شود. که برای نمونه‌های مثبت و منفی ماتریس‌های قطری $Y_1 = diag(\widetilde{y_1}, \dots, \widetilde{y_p})$ و $Y_2 = diag(\widehat{y_1}, \dots, \widehat{y_q})$ تعریف می‌شود. در اینجا $\widetilde{y_i} = 2\widetilde{m_i} - 1, (i = 1, \dots, p)$ و $\widehat{y_j} = 2\widehat{m_j} - 1, (j = 1, \dots, q)$ می‌باشد. زمانی $m_i = 1$ می‌باشد که x_i نمونه مثبت باشد و برچسب مرتبط به آن $Y_i = 2m_i - 1 = 1$ می‌باشد. زمانی $m_i = 0$ می‌باشد که x_i نمونه منفی باشد و برچسب مرتبط به آن $Y_i = 2m_i - 1 = -1$ می‌باشد. در این رساله کلاس

مثبت نشان‌دهنده شی و کلاس منفی نشان‌دهنده پس‌زمینه می‌باشد. مسئله بهینه‌سازی به صورت رابطه‌های (۳-۲۰) و (۳-۲۱) تعریف می‌شود:

$$\min_{w_1,b_1,\xi_2} \frac{1}{2}\|Aw_1 + e_1b_1\|_2^2 + \frac{1}{2}c_1(w_1^2 + b_1^2) + c_2e_2^T\xi_2$$
$$s.t. \quad Y_2(Bw_1 + e_2b_1) \geq Y_2^2e_2 - Y_2^2\xi_2, \xi_2 \geq 0 \qquad (۲۰\text{-}۳)$$

$$\min_{w_2,b_2,\xi_1} \frac{1}{2}\|Bw_2 + e_2b_2\|_2^2 + \frac{1}{2}c_3(w_2^2 + b_2^2) + c_4e_1^T\xi_1$$
$$s.t. \quad Y_1(Aw_2 + e_1b_2) \geq Y_1^2e_1 - Y_1^2\xi_1, \xi_1 \geq 0 \qquad (۲۱\text{-}۳)$$

c_1،c_2،c_3 و c_4 پارامترهای جریمه مثبت، ξ_1و ξ_2 متغیرهای نرم، ماتریس $A\epsilon\Re^{P\times n}$ نمونه‌های مرتبط به کلاس مثبت و $B\epsilon\Re^{q\times n}$ نمونه‌های مرتبط به کلاس منفی می‌باشند. بعد از حل این رابطه توسط لاگرانژ چندگانه که در مرجع [۱۰۰] به تفصیل بیان شد رابطه‌های (۳-۲۲) و (۳-۲۳) به دست می‌آیند.

$$v_1 = (H^TH + c_1I)^{-1}G^TT_2^2\alpha \quad where\ v_1 = [w_1^T b_1]^T, H = [Be_2] \qquad (۲۲\text{-}۳)$$

$$v_2 = (G^TG + c_3I)^{-1}H^TY_1^2\gamma \quad where\ v_2 = [w_2^T b_2]^T, G = [H\ Ae_1] \qquad (۲۳\text{-}۳)$$

بعد از به دست آوردن v_1 و v_2، برای هر نمونه جدید $x \in \Re^n$ برچسب نمونه توسط رابطه (۳-۲۴) به دست می‌آید.

$$x \in W_k, k = \underset{i=1,2}{\operatorname{argmin}} \left\{ \frac{|w_1^Tx + b_1|}{\|w_1\|}, \frac{|w_2^Tx + b_2|}{\|w_2\|} \right\} \qquad (۲۴\text{-}۳)$$

برای محاسبه مقدار عضویت از تابع عضویت معرفی شده در مرجع [۱۰۱] استفاده شد. برای نمونه x_i با برچسب مثبت (۱+) عضویت فازی به صورت رابطه (۳-۲۵) بیان می‌شود:

$$m_1(x_i) = 0.5 + \frac{\exp(C_0(d_{-1}(x_i) - d_1(x_i))/d) - \exp(-C_0)}{2(\exp(C_0) - \exp(-C_0))} \qquad (۲۵\text{-}۳)$$
$$m_{-1}(x_i) = 1 - m_1(x_i)$$

برای نمونه x_i با برچسب منفی (۱-) عضویت فازی به صورت رابطه (۳-۲۶) بیان می‌شود.

$$m_{-1}(x_i) = 0.5 + \frac{\exp(C_0(d_1(x_i) - d_{-1}(x_i))/d) - \exp(-C_0)}{2(\exp(C_0) - \exp(-C_0))}$$
$$m_1(x_i) = 1 - m_{-1}(x_i) \qquad (۲۶-۳)$$

$d_1(x_i)$ فاصله بین x_i و میانگین کلاس مثبت، $d_{-1}(x_i)$ فاصله بین x_i و میانگین کلاس منفی، d فاصله بین میانگین کلاس‌های مثبت و منفی، c_0 یک ثابت به عنوان کنترل‌کننده تابع عضویت می‌باشد.

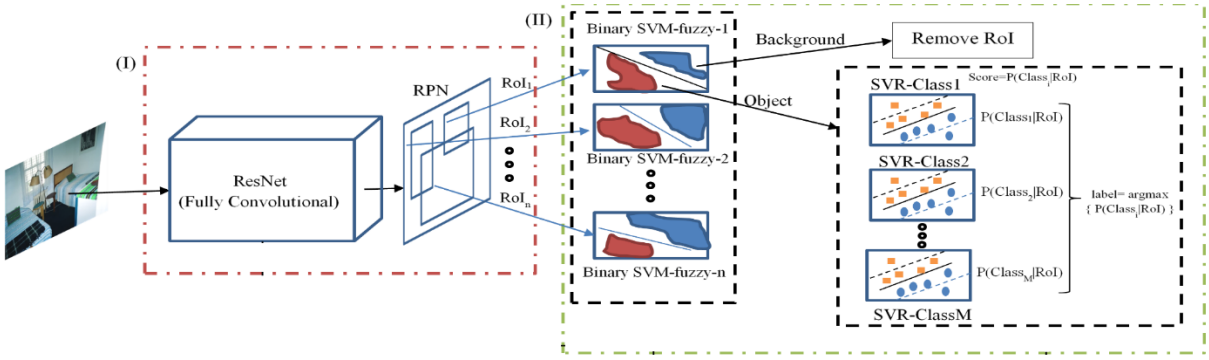
۳ - ۵ - ۴ استخراج برچسب و نواحی اشیاء (نوآوری)

در R-FCN پس‌زمینه به عنوان یک کلاس مجزا در نظر گرفته می‌شود. در روش پیشنهادی نمی‌خواهیم تعداد کلاس‌ها را افزایش دهیم در همین راستا یک فاز مجزا برای جدا نمودن پس‌زمینه از اشیاء در نظر گرفته می‌شود. در همین راستا ابتدا یک سیستم فازی دوکلاسه مبتنی بر SVM طراحی شد که مشخص می‌کند آیا ناحیه مربوط به شی می‌باشد یا نه، در صورتی که متعلق به شی نباشد برچسب‌زنی برای این ناحیه انجام نمی‌شود. معمولاً برای هر تصویر تعداد زیادی ناحیه کاندیدا معرفی می‌شود که با این عمل فقط نواحی که می‌توانند به عنوان شی باشند مورد ارزیابی قرار می‌گیرند. این عمل زمان برچسب‌زنی برای کل تصویر را بسیار کاهش می‌دهد.

روش R-FCN از رای‌دهی برای انتخاب برچسب هر RoI استفاده می‌کند. ولی در روش پیشنهادی از چند SVR به صورت موازی استفاده می‌شود. در واقع به ازای هر کلاس از اشیاء، یک SVR در نظر گرفته می‌شود. در روش پیشنهادی هر SVR میزان تعلق هر RoI به هر کلاس را مشخص می‌کند. در نهایت بین تمام RoI متعلق به شی، عمل ماکزیمم‌گیری انجام می‌شود و برچسب هر RoI مشخص می‌شود. در روش پیشنهادی همه‌ی SVRها به صورت خط لوله اجرا می‌شوند و زمان پردازش کاهش می‌یابد. در این روش برای هر شی یک SVR اجرا می‌شود که برای این منظور هر SVR در یک خط لوله مجزا اجرا می‌شود. شمای کلی روش پیشنهادی در شکل ۳-۵ نشان داده می‌شود. در واقع در هر

SVRمیزان تعلق هر RoI به یک کلاس مشخص محاسبه می‌شود. در نهایت با استفاده از رابطه (۳-۲۷) برچسب کلاس تعیین می‌شود. در این رابطه n تعداد کلاس‌ها می‌باشد.

$$ClassNumber = \underset{i}{\overset{n}{arg\,Max}}\{P(Class_i|RoI)\} \qquad (۳-۲۷)$$

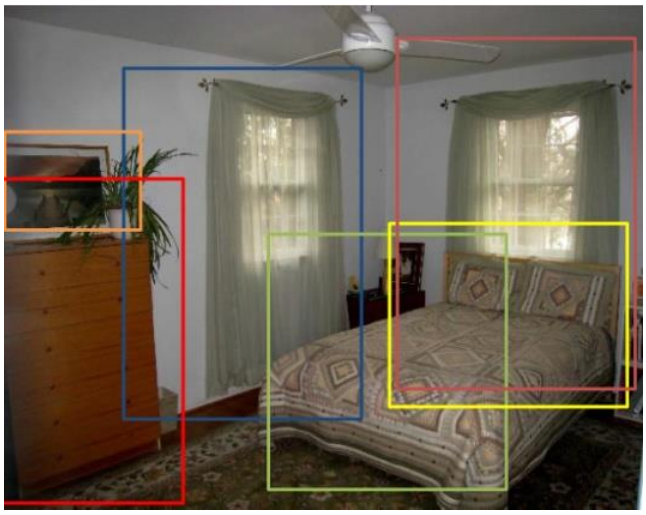


شکل ۳-۵ روش پیشنهادی برای استخراج نواحی کاندیدا و برچسب‌زنی تصویر

۳ - ۵ - ۵ استخراج مجموعه کلمات بصری (*نوآوری*)

همان طور که در قسمت‌های قبل اشاره شد، به دست آوردن ویژگی‌های معنایی سطح بالا می‌تواند در بالا بردن دقت سیستم دسته‌بندی صحنه تصویر بسیار موثر باشد. در روش پیشنهادی برای به دست آوردن نواحی و برچسب کلمات در تصویر از شبکه عصبی عمیق استفاده می‌شود. نتایج بدست آمده از این مرحله نشان می‌دهدکه بسیاری از نواحی مرتبط به اشیاء به درستی انتخاب نمی‌شوند و بسیاری از نواحی همپوشانی مناسبی با اشیاء موجود در تصویر ندارند. برای نمونه شکل ۳-۶ تعدادی از نواحی RoI انتخاب شده توسط این روش را نشان می‌دهد. در تصویر نمونه، بعضی از نواحی پیشنهادی دارای مشکلاتی مانند همپوشانی چندشی و همپوشانی شی با پس‌زمینه می‌باشند. در واقع این نواحی به تنهایی نمی‌توانند برای دسته‌بندی صحنه تصویر مناسب باشند. در این بخش از این نواحی برای استخراج ویژگی‌های بصری سطح بالا مانند مجموعه کلمات بصری استفاده می‌شود.

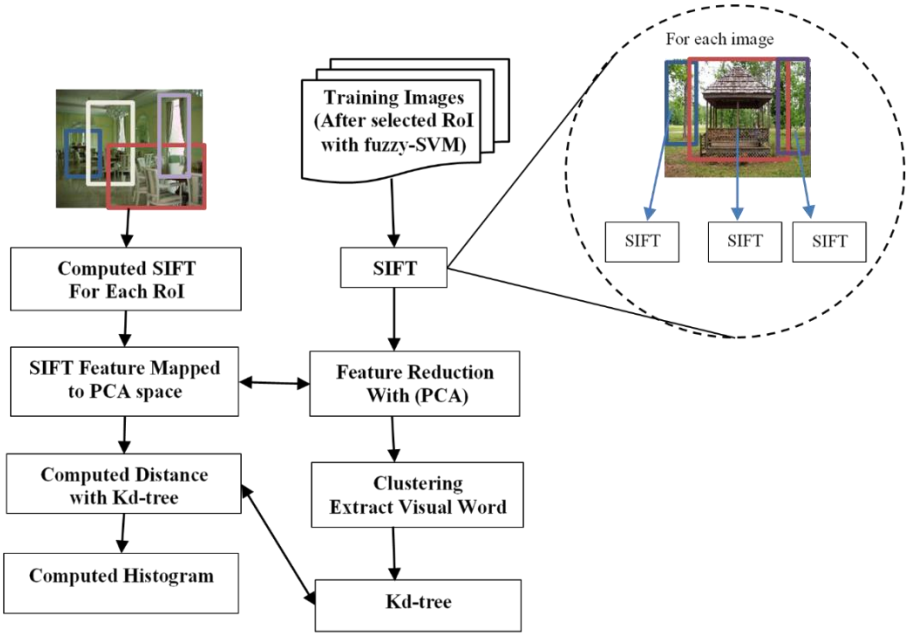
روش مجموعه کلمات بصری به صورت خط لوله در سال‌های اخیر عملکرد خوبی در استخراج ویژگی‌های سطح بالا و دسته‌بندی در تصویر از خود نشان داده است. این روش عملکرد خوبی در شرایطی با تغییر شدت روشنایی و نویز دارد [۱۰۲]. این روش شامل مراحل استخراج ویژگی، پیش‌پردازش ویژگی، کتابچه کد، کدگذاری ویژگی، انتخاب و نرمال کردن می‌باشد.



شکل ۳-۶ نمونه‌ای از خروجی شبکه پیشنهادی رساله در محاسبه نواحی کاندیدا

در مرحله استخراج ویژگی از ویژگی‌های محلی سطح پایین استفاده می‌شود. در مرجع [۱۰۳] اشاره شد که استخراج ویژگی‌های محلی شامل شناسایی نواحی محلی و توصیف این نواحی انتخابی می‌باشند. در تحقیقات انجام شده در سال‌های اخیر، برای استخراج نواحی محلی از کل تصویر استفاده شده است. ولی در این رساله، از نواحی مرتبط با مرحله قبل استفاده می‌شود. در واقع در این روش نواحی محلی از نواحی مرتبط با شی استخراج می‌شوند. همچنین ویژگی‌های سطح پایین توسط روش SIFT در نواحی محلی مرتبط با اشیاء استخراج می‌شوند. این ویژگی برای هر RoI به صورت مجزا محاسبه می‌شود. توصیفگرهای محلی سطح پایین معمولاً داری بعد بالا و همبستگی قوی می‌باشند [۱۰۲]. در همین راستا، برای کاهش بعد از روش PCA استفاده می‌شود. بعد از کاهش ویژگی برای تولید کتابچه کد معمولا دو روش کلی وجود دارد. در روش اول، فضای ویژگی به چندین ناحیه تقسیم می‌شود. هر ناحیه توسط یک مرکز نمایش داده می‌شود. در روش دوم، یک مدل برای نمایش توزیع احتمال ویژگی‌ها تولید می‌شود. در روش پیشنهادی برای بالا بردن سرعت از روش دوم استفاده می‌شود. در مرحله بعد باید

کدگذاری ویژگی‌ها انجام شود که شامل روش‌هایی مانند رای‌گیری، بازسازی بر اساس روش کدگذاری، بردار پشتیبان بر اساس روش کدگذاری می‌باشد. در این مرحله، میزان نزدیکی هر نمونه به مرکز خوشه در نظر گرفته می‌شود. در مرحله بعدی که مرحله نرمال‌سازی و انتخاب می‌باشد، برداز ویژگی نهایی توسط SumPooling و MaxPooling برای هر تصویر استفاده می‌شود. روش پیشنهادی برای انتخاب BoVW در شکل ۳-۷ نشان داده شد.



شکل ۳-۷ مراحل استفاده شده برای تولید BoVM در رساله

همانطور که در شکل ۳-۷ مشاهده می‌شود، در مرحله آموزش، RoIهای شناسایی شده توسط SVM فازی به عنوان نواحی مرتبط با اشیاء برای هر تصویر استفاده می‌شوند. در مرحله بعد از روش SIFT برای استخراج ویژگی و همچنین برای کاهش ویژگی از روش PCA استفاده می‌شود. روش‌های خوشه‌بندی بسیاری مانند k-means، خوشه‌بندی سلسله مراتبی و خوشه‌بندی طیفی[۱] برای محاسبه کتابچه کد وجود دارند. روش k-means به دلیل سادگی در پیاده‌سازی برای ساختن کتابچه کد بسیار رایج می‌باشد. در روش پیشنهادی، از یک k-means توسعه یافته براساس مرجع [۱۰۴] استفاده شده است. دلیل استفاده از این روش سریعتر بودن این روش نسبت به k-means معمولی است [۱۰۴].

[۱] Spectral

تاکنون کلمات بصری در مجموعه آموزش استخراج شده است. این کلمات برای مرحله آزمون استفاده می‌شوند. در تصویر ورودی آموزش، بعد از استخراج ویژگی و نگاشت آنها در فضای PCA در مرحله کدگذاری از روش kd-tree (درخت k- بعدی) برای محاسبه فاصله بین هر نمونه تا مرکز خوشه استفاده می‌شود. استفاده از این روش پیچیدگی زمانی را از $O(n \times m)$ به $O(m \times \log n)$ کاهش می‌دهد. در اینجا m و n به ترتیب تعداد خوشه‌ها و تعداد نمونه‌ها می‌باشد. در نهایت از MaxPooling برای محاسبه تعداد تکرار هر نمونه در خوشه‌ها استفاده می‌شود و برای هر تصویر تعداد تکرار هر کلمه بصری (هیستوگرام) محاسبه می‌شود.

باید به این نکته اشاره نمود که در روش پیشنهادی می‌توان در فاز خوشه‌بندی برچسب هر ناحیه را در نظر گرفت. در واقع خوشه‌بندی را با ناحیه‌هایی با برچسب مشابه انجام داد. ولی با توجه به آزمایشات انجام شده، در نظر گرفتن برچسب نواحی عملکرد سیستم نهایی را کاهش می‌دهد. لازم به ذکر است که در مرحله دسته‌بندی از برچسب اشیاء برای هر کلمه بصری استفاده می‌شود

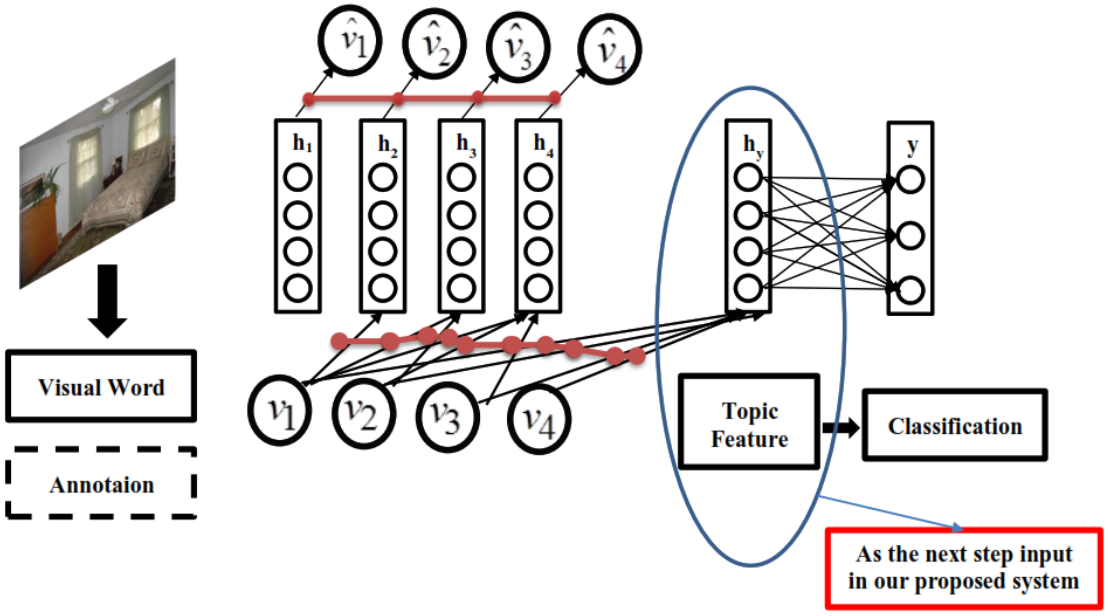
۳ - ۵ - ۶ استخراج مدل موضوعی

در سال‌های اخیر روش‌های بسیاری برای استخراج موضوعات نهفته در تصاویر معرفی شده است. در این روش‌ها از داده‌های چندوجهی[۱] برای مدل سازی استفاده می‌شود. یکی از معروفترین این روش‌ها روش LDA می‌باشد. این روش در ابتدا برای مدل‌سازی اسناد استفاده شد. روش‌های بسیاری مانند Corr-LDA، LDA چندوجهی و MDRF در راستای توسعه LDA معرفی شد. آزمایشات انجام شده نشان دهنده پیچیدگی زمانی بالای این روش‌ها در مجموعه داده‌های حجیم می‌باشد. در مرجع [۱] روشی با عنوان تخمین‌گر توزیع شده خودرگرسیو عصبی اسناد برای مدل‌سازی موضوع معرفی شد که این شبکه برپایه معماری عمیق می‌باشد و عملکرد مناسبی در زمان اجرا داشته است. از آنجایی که در آزمایشات انجام شده در این رساله، این روش از لحاظ پیچیدگی زمانی و دقت بسیار قدرتمند است، از این روش

[۱] Multimodal

برای محاسبه موضوعات پنهان استفاده شده است. نمایش کلی از روش پیشنهادی در مرجع [۱] در شکل ۳-۸ نشان داده شده است. این روش یک معماری نوین با نظارت از روش DocNADE می‌باشد. این معماری قدرت یادگیری ویژگی موضوعی را افزایش داده و همچنین می‌توان از این روش برای اشتراک اطلاعات مرتبط با کلمات بصری، برچسب نواحی موجود در تصویر و برچسب‌های کلاس (برچسب صحنه تصویر) استفاده کرد. در سیستم استفاده شده در مرجع [۱]، برچسب نواحی موجود در تصویر از قبل به عنوان ورودی آماده بوده است. در واقع این روش برای مجموعه داده‌هایی استفاده می‌شود که برچسب‌ها از قبل وجود داشته باشند. ولی در این رساله، برچسب اشیاء به صورت خودکار توسط روش پیشنهادی در مرحله قبل محاسبه و به عنوان ورودی این روش در نظر گرفته می‌شوند.

روش ارائه شده بر پایه یک شبکه عمیق برای محاسبه برچسب صجنه تصویر می‌باشد. لایه آخر، برچسب هر صحنه را مشخص می‌کند. لایه ماقبل آخر، ویژگی‌های موضوعی نهفته در تصویر را محاسبه می‌کند. در روش پیشنهادی از خروجی لایه ماقبل آخر برای استخراج ویژگی موضوعی برای مرجله بعد استفاده می‌شود. در واقع در روش پیشنهادی برای دسته‌بندی صحنه از روش مرجع [۱] استفاده نمی‌شود و فقط از خروجی لایه ماقبل آخر استفاده می‌شود. باید به این نکته اشاره نمود که محل قرار گرفتن هر شی در تصویر ویژگی مهم و با ارزشی می‌باشد. در اینجا هر تصویر به ناحیه‌های مساوی تقسیم و محل قرار گرفتن هر ناحیه بصری محاسبه می‌شود.



شکل ۳-۸ روش پیشنهادی در مرجع [۱] برای استخراج موضوعات پنهان. خروجی لایه استخراج ویژگی‌های موضوعی به عنوان ورودی مرحله بعد روش پیشنهادی می‌باشد.

۳ - ۵ - ۷ دسته‌بندی صحنه تصویر بر پایه ویژگی‌های استخراج شده

در مرحله بعد از روش بیزین ساده برای دسته‌بندی صحنه تصویر استفاده می‌شود. چندین نوع ویژگی به عنوان ورودی این دسته‌بند درنظر گرفته شد. ویژگی‌های استفاده شده در جدول ۳-۱ نشان داده شده است. با توجه به جدول ۳-۱، ویژگی‌های استخراج شده از شبکه عمیق ResNet، موضوعات نهفته درتصویر و برچسب استخراج شده از اشیاء برای کلمات بصری، به عنوان ویژگی‌های استفاده شده برای دسته‌بند بیزین در نظر گرفته شده است. روش PCA برای کاهش ویژگی‌های استخراج شده از شبکه عمیق ResNet استفاده می‌شود.

جدول ۳-۱ ویژگی استخراج شده برای کلاس‌بند بیزین

	روش ها		
	ResNet	BoVM+SupDocNADE	SVR
ویژگی‌ها	ویژگی آخرین لایه پیچش برای اشیاء با اعمال PCA	ویژگی موضوعات نهفته	برچسب هر کلمه بصری

۳ - ۶ جمع‌بندی

در این فصل روش‌های مطرح شده برای دسته‌بندی صحنه‌های تصویر پیچیده معرفی شده است. روش پیشنهادی در این بخش برپایه روش‌های پایه یادگیری ماشین و روش‌های استخراج مدل‌های موضوعی و یادگیری عمیق می‌باشد. در روش ارائه شده بر پایه روش‌های سنتی یادگیری ماشین، از روش ترکیبی بر پایه Boosting، نواحی کاندیدا و پیچش برپایه SVD استفاده می‌شود. روش دوم بر پایه شبکه عصبی عمیق تمام پیچش ResNet، یک معماری نوین برای استخراج نواحی و برچسب نواحی و همچنین مدل‌های موضوعی بوده است. در فصل آتی مراحل مختلف پیاده سازی تشریح می‌شود.

۴ پیاده‌سازی و نتایج آزمایشات

۴ - ۱ مقدمه

در این فصل مراحل پیاده‌سازی و تحلیل نتایج روش‌های پیشنهادی ارائه می‌شود. همان‌طور که در فصل قبل اشاره شد، دو نوع مجموعه داده از لحاظ پیچیدگی صحنه تصویر در نظر گرفته شده است. در همین راستا در گام نخست پیاده‌سازی و تحلیل نتایج بر روی صحنه‌هایی با پیچیدگی کم (تصاویر نامتعارف دو کلاسه) ارائه می‌شود. در قسمت دوم این فصل، مراحل پیاده‌سازی دو روش ارائه شده در مجموعه داده‌هایی با پیچیدگی زیاد معرفی می‌شوند و نتایج این روش‌ها تحلیل می‌گردد.

۴ - ۲ پیاده‌سازی و نتایج دسته‌بندی تصاویر نامتعارف

در این بخش، به بیان مجموعه داده و مقایسه روش پیشنهادی با برخی از روش‌های انجام شده در شناسایی تصاویر نامتعارف پرداخته می‌شود. لازم به ذکر است که پیاده‌سازی‌های انجام شده در زبان برنامه‌نویسی متلب بر روی سخت‌افزاری با مشخصات حافظه 4GB و پردازنده Core i7 M620 2.67GHz می‌باشد. لازم به ذکر است، تمامی زمان‌های گزارش شده در مرحله آموزش و آزمون در یک سیستم عامل یکسان، سخت افزار و محیط برنامه نویسی یکسان انجام شده است. نتایج توسط توابع یکسان در محیط متلب مشابه کارهای انجام شده در سال‌های اخیر محاسبه شده است. زمان مرحله آزمون، متوسط زمان به ازای ۱۰۰ تکرار برای کل تصاویر می‌باشد.

۴ - ۲ - ۱ مجموعه داده تصاویر نامتعارف

در دسته‌بندی تصاویر نامتعارف از دو مجموعه داده استفاده شده است. در ادامه به معرفی این مجموعه داده‌ها پرداخته می‌شود.

مجموعه‌داده اول استفاده شده در این تحقیق از مرجع [۱۰۵] به دست آمده است. تصاویر استفاده شده در این مجموعه، شامل دو دسته تصویر متعارف و نامتعارف هستند. در این مجموعه داده از ۱۸۳۳۵ تصویر جمع آوری شده از صفحات مختلف وب، ۹۰۵۹ تصویر متعارف و ۹۲۹۵ تصویر نامتعارف می‌باشد که در تصاویر انتخاب شده بسیاری از حالت‌های ممکن و قسمت‌های مختلف از هر دو دسته در نظر گرفته شده است.

در مجموعه داده اول استفاده شده، چگالی توزیع تصاویر، متعادل است. اما در وب، توزیع تصاویر نامتعارف و متعارف یکسان نیستند و تعداد تصاویر متعارف بسیار بیشتر است. به همین دلیل مجموعه داده مورد نظر از اینترنت جمع‌آوری شده است. در مجموعه داده دوم از ۳۵۶۰۰ تصویر استفاده شد که ۲۶۷۰۰ نمونه از آن‌ها، تصاویر متعارف و ۸۹۰۰ نمونه باقیمانده، تصاویر نامتعارف هستند.

۴ - ۲ - ۲ معیار ارزیابی برای دسته‌بندی تصاویر نامتعارف

برای نشان دادن عملکرد روش پیشنهادی در شناسایی تصاویر نامتعارف از معیارهای کمی دقت[۱](PR)، یادآوری[۲](RE) و F1 استفاده می‌شود که این معیارها در مقاله‌های مختلفی برای ارزیابی روش‌های تشخیص تصاویر نامتعارف استفاده شده‌اند که در رابطه (۴-۱) تا رابطه (۴-۳) نشان داده شده است.

$$PR = \frac{TP}{TP + FP} \qquad (۴-۱)$$

$$RE = \frac{TP}{TP + FN} \qquad (۴-۲)$$

$$F1 = \frac{2 * RE * PR}{PR + RE} \qquad (۴-۳)$$

در رابطه بالا منظور از TP تعداد تصاویری است که به درستی توسط طبقه‌بند به طبقه C_i انتساب یافته و FN یعنی تعداد تصاویر متعلق به سایر طبقه‌ها که به اشتباه توسط طبقه‌بند به طبقه C_i منسوب شده‌اند. FP تعداد تصاویر متعلق به طبقه C_i که توسط طبقه‌بند به سایر طبقه‌ها تعلق گرفته است. معیار

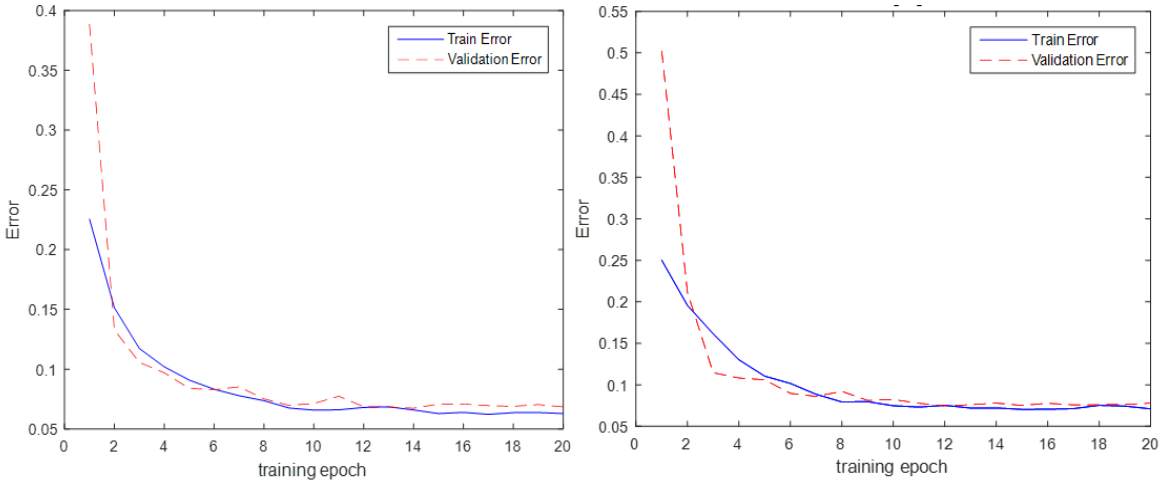
[۱] Precision
[۲] Recall

TN تعداد تصاویری است که متعلق به طبقه C_i نبوده‌اند و توسط طبقه‌بند به این طبقه اختصاص داده نشده‌اند.

۴ - ۲ - ۳ نتایج آزمایش‌ها

در مجموعه داده‌ی استفاده شده دوسوم داده‌ها برای آموزش و یک‌سوم برای آزمون در نظر گرفته شده است. ساختار کلی شبکه عمیق پیشنهادی در فصل قبل معرفی گردید. در مجموعه‌ی آموزش یک‌سوم برای اعتبارسنجی در نظر گرفته می‌شود.

در گام نخست، مقایسه‌ای بین معماری معروف AlexNet و معماری پیشنهادی انجام شده است. شکل ۴-۱ نتایج بدست آمده از دسته‌بندی در مرحله آموزش و اعتبارسنجی، با دو معماری شبکه عصبی پیچش پیشنهادی و مرجع [۸۷] را نشان می‌دهد. همانطور که از شکل ۴-۱ مشخص است، شبکه پیچش پیشنهادی در مجموعه آموزش عملکرد بهتری داشته است. معماری پیشنهادی میزان خطای کمتری در شبکه پیچش نسبت به روش مرجع [۸۷] در مجموعه آموزش و اعتبارسنجی از خود نشان داده است.



ب) الف)

شکل ۴-۱ نتایج مرحله آموزش و اعتبارسنجی در شبکه عصبی پیچش (خط پر و خط‌چین به ترتیب نشان‌دهنده مقدار خطای مجموعه آموزش و اعتبارسنجی است). الف) نتایج با معماری شبکه مرجع [۸۷] ب) نتایج با معماری شبکه پیشنهادی

جدول ۴-۱ مقایسه‌ی نتایج دو معماری در مجموعه آزمون را نشان می‌دهد. برای مقایسه از معیار درصد TP و FP استفاده شد. درصد TP نشان دهنده نسبت تعداد تصاویر نامتعارف صحیح شناسایی شده به

کل تصاویر نامتعارف و درصد FP نشان‌دهنده‌ی نسبت تعداد تصاویر متعارف ناصحیح شناسایی شده به کل تصاویر متعارف تعریف شده است. با توجه به جدول ۴-۱ نتایج به دست آمده از معماری پیشنهادی نسبت به معماری AlexNet در مرحله آزمون عملکرد بهتری دارد.

جدول ۴-۱ نتایج بدست آمده به صورت تفکیکی توسط روش پیشنهادی و شبکه‌عصبی پیچش [۸۷] خروجی پیش‌بینی شده

	نامتعارف(مثبت)		متعارف(منفی)	
	پیشنهادی	[۸۷]	پیشنهادی	[۸۷]
نامتعارف (مثبت)	**TP**		**FN**	
	٪۹۱,۳۵	٪۸۹,۵۰	**٪۸,۶۵**	٪۱۰,۵
متعارف (منفی)	**FP**		**TN**	
	٪۵,۱۱	٪۹,۲۱	**٪۹۴,۸۹**	٪۹۰,۷۹

جدول ۴-۲ معیارهای معرفی شده در بخش ارزیابی را برروی این دو معماری نشان می‌دهد. نتایج جدول ۴-۲ نشان می‌دهد که شبکه پیچش با معماری پیشنهادی عملکرد بهتری نسبت به معماری پیشنهاد شده در مرجع [۸۷] دارد. در جدول ۴-۲، زمان شبکه در مرحله آموزش و آزمایش برای هر تصویر ارائه شده است، مقایسه زمان اجرای الگوریتم‌ها، بیانگر بهبود عملکرد زمانی شبکه پیچش پیشنهادی نسبت به روش مرجع [۸۷] در مرحله آموزش و آزمایش است.

جدول ۴-۲ نتایج به دست آمده از مقایسه شبکه‌های عصبی پیچش با معماری متفاوت

معیار / روش	PR	RE	F1	زمان- برای هر تصویر	
				آموزش	آزمایش
روش پیشنهادی	**٪۹۴,۷۰**	**٪۹۱,۳۵**	**٪۹۲,۹۹**	**۰,۵۷**	**۰,۰۷۶**
روش مرجع [۸۷]	٪۹۰,۶۶	٪۸۹,۹۵	٪۹۰,۳۰	۰,۶۱	۰,۰۸۸

در جدول ۴-۳ مقایسه‌ی بین روش پیشنهادی و برخی روش‌های تشخیص تصاویر نامتعارف در پایگاه داده اول را نشان می‌دهد. با توجه به نتایج بدست آمده در جدول ۴-۳، روش یادگیری عمیق دارای عملکرد مناسبی است و شبکه عصبی پیچش با معماری پیشنهادی، بهترین عملکرد را از خود نشان می‌دهد. در فیلترینگ هوشمند تصاویر، علاوه‌بر تشخیص صحیح تصاویر نامتعارف، تشخیص صحیح

تصاویر متعارف بسیار با اهمیت است که روش پیشنهادی در هر دو قسمت، بهتر از همه‌ی روش‌های مطرح شده عمل می‌کند. همان‌طور که در فصل پیش بیان شد روش‌هایی مختلفی در چهار دسته برای شناسایی تصاویر نامتعارف وجود دارند که روش‌های انتخابی برای مقایسه، همه دسته‌ها را شامل می-شوند.

در مرجع [۸۷] از شبکه‌عصبی پیچش با معماری AlexNet برای دسته‌بندی تصاویر مجموعه داده ImageNet استفاده گردید که در این تحقیق برای شناسایی تصاویر نامتعارف به کار برده شده است. از آنجایی که تاکنون پژوهشی برای شناسایی تصاویر نامتعارف توسط شبکه‌عصبی پیچش AlexNet صورت نگرفته، پیاده‌سازی و نتایج این روش توسط نویسنده این تحقیق انجام شده است.

جدول ۴-۳ نتایج به دست آمده از روش‌های مختلف برای طبقه‌بندی تصاویر نامتعارف

روش / معیار ارزیابی	روش پیشنهادی	مرجع [۸۷]	مرجع [۱۰۵]	مرجع [۱۰۶]	مرجع [۱۰۷]	مرجع [۶۳]
TP	**٪۹۱,۳۵**	٪۸۹,۵	٪۸۶,۹۷	٪۸۳,۶۵	٪۷۸,۳۷	٪۹۰,۹۳
FP	**٪۵,۱۱**	٪۹,۲۱	٪۵,۳۸	٪۶,۶۴	٪۱۶,۹۸	٪۱۸,۹۴
FN	**٪۸,۶۵**	٪۱۰,۵	٪۱۳,۰۳	٪۱۶,۳۵	٪۲۱,۶۳	٪۹,۰۷
TN	**٪۹۴,۸۹**	٪۹۰,۷۹	٪۹۴,۶۲	٪۹۳,۳۶	٪۸۳,۰۲	٪۸۱,۰۶

شایان ذکر است که نتایج جدول ۴-۳، از مرجع [۱۰۵] گرفته شده است. این نتایج بر روی پایگاه داده یکسان، گزارش شده است. مرجع [۱۰۶] روشی هوشمند بر پایه بافت و ویژگی‌های بصری معرفی کرده که این ویژگی‌های بصری بر پایه‌ی شکل نواحی مرتبط به پوست بدن استخراج شده است. در نهایت تصاویر نامتعارف بر پایه ویژگی‌های معرفی شده توسط یک طبقه‌بند با ساختار سلسله‌مراتبی شناسایی شده‌اند. مرجع [۱۰۷] نیز از ویژگی‌های شکل نواحی پوست افراد استفاده کرده است. این مقاله یک سیستم طبقه‌بند با چندین بیز[1] برای شناسایی نواحی پوست معرفی کرد. مرجع [۶۳] از روش‌های

[1] Multi-Bayes

بازیابی تصاویر برپایه محتوا استفاده کرده است. در این روش ابتدا پس‌زمینه حذف شده و نواحی مرتبط با پیسکل‌های پوست استخراج می‌گردد. سپس تصویر نامتعارف توسط ویژگی‌های رنگ، بافت و شکل با کمک معیار شباهت تشخیص داده شد.

همانطور که در بخش معرفی مجموعه داده ذکر شد، چگالی توزیع تصاویر در مجموعه داده اول متعادل است. اما در وب، توزیع تصاویر نامتعارف و متعارف یکسان نیستند و تعداد تصاویر متعارف بسیار بیشتر است. به همین دلیل روش پیشنهادی بر روی مجموعه داده دوم مورد ارزیابی قرار می‌گیرد. نتایج به دست آمده برروی این مجموعه داده، در جدول ۴-۴ نشان داده شده است. با توجه به نتایج به دست آمده در این آزمایش، شبکه عصبی با معماری پیشنهادی، عملکرد بهتری بر روی مجموعه‌داده با چگالی توزیع نامتعادل، دارد.

جدول ۴-۴ نتایج بدست آمده از مقایسه شبکه‌های عصبی پیچش با معماری متفاوت

معیار ارزیابی / روش	PR	RE	F1
روش پیشنهادی	۹۲,۶۰	۸۷,۳۵	۸۹,۷۹
روش مرجع [۸۷]	**۸۷,۶۶**	**۷۹,۹۵**	**۸۳,۶۲**

۴ - ۳ پیاده سازی و نتایج دسته‌بندی صحنه‌های پیچیده

در این بخش، پیاده‌سازی‌ها و نتایج به دست آمده توسط دو روش پیشنهادی در صحنه‌هایی با پیچیدگی زیاد معرفی می‌شوند. لازم به ذکر است که روش پیشنهادی بسیار وابسته به استخراج اشیاء می‌باشد در همین راستا پیاده‌سازی‌ها و نتایج روش پیشنهادی برای استخراج اشیاء موجود در تصویر نیز معرفی می‌شوند. لازم به ذکر است که پیاده‌سازی‌های انجام شده در زبان برنامه‌نویسی پایتون بر روی سخت‌افزاری با مشخصات،کارت گرافیک GTX 1080، حافظه ۳۲GB و پردازنده Core i7 4790k

4GHz می‌باشد. در پیاده سازی‌های انجام شده مرتبط به یادگیری عمیق از چارچوب Caffe استفاده شده است. برای تغییر اندازه تصاویر از روش درون یابی دوخطی[1] استفاده شده است.

لازم به ذکر است، تمامی زمان‌های گزارش شده در مرحله آموزش و آزمون در یک سیستم عامل یکسان، سخت افزار و محیط برنامه نویسی یکسان انجام شده است. نتایج توسط توابع یکسان در محیط پایتون مشابه کارهای انجام شده در سال‌های اخیر محاسبه شده است. زمان مرحله آزمون، متوسط زمان به ازای ۱۰۰ تکرار برای کل تصاویر می‌باشد.

۴ - ۴ مجموعه داده صحنه های پیچیده

در این بخش، چهار پایگاه‌داده استاندارد در دسته‌بندی صحنه تصویر معرفی می‌شوند. پایگاه داده‌های استفاده شده در این رساله شامل SUN، MIT-67، UIUC Sport و Scence_15 می‌باشند.

۴ - ۴ - ۱ مجموعه داده SUN

در این رساله برای آشکارسازی و برچسب‌زنی تصاویر از پایگاه داده SUN[2] [۱۰۸] استفاده شده است. این پایگاه داده یک مجموعه بزرگ از تصاویر برای دسته‌بندی صحنه‌ی تصویر است. این پایگاه داده، شامل ۱۳۱۰۶۷ تصویر در ۹۰۸ دسته مختلف و شامل ۳۸۱۹ شی متفاوت در تصاویر است. تصاویر به دست آمده در حالت‌های مختلف مانند محیط بسته، باز، طبیعی و غیره تهیه گردید. از آنجایی که هدف در این بخش، معرفی پایگاه داده‌ای برای آشکارسازی و برچسب‌زنی اشیاء می‌باشد که اشیایی انتخابی از نظر تکرار قابل قبول بوده‌اند. در واقع اشیایی که از تعداد تکرار پایینی برخوردارند، حذف شدند. در همین راستا، ۲۱۵۱۸ تصویر در پنج دسته مختلف در شرایط متفاوت انتخاب گردید. مجموعه داده انتخابی شامل ۳۶ شی می‌باشد. در جدول ۴-۵ تعداد تصاویر هر شی نشان داده شده است. در هر تصویر امکان

[1] Bilinear Interpolation

[2] Scene Understanding

تکرار یکی شی وجود دارد. در این مجموعه داده، ۷۰٪ از داده‌ها برای آموزش واعتبارسنجی و ۳۰٪ برای آزمون در نظر گرفته شده‌اند.

جدول ۴-۵ اشیاء موجود پایگاه داده SUN

شی	**Countertop**	**Buildings**	**Chair**	**Ceiling**	**Cabinet**	**chair occluded**
تعداد تصاویر	۱۰۹	۴۵۸	۳۴۳	۱۱۱۵	۱۳۲۵	۱۰۵۷
شی	**night table**	**Mirror**	**microwave**	**Flowers**	**Floor**	**Faucet**
تعداد تصاویر	۲۰۶	۳۲۹	۱۴۹	۱۰۹	۱۵۶۶	۵۵۳
شی	**Stove**	**Skyscraper**	**Sky**	**Sink**	**Plant**	**pillow occluded**
تعداد تصاویر	۲۳۱	۶۲۲	۲۴۵	۲۵۵	۴۵۱	۳۱۴
شی	**Towel**	**Toilet**	**Table**	**Painting**	**Pillow**	**night table occluded**
تعداد تصاویر	۳۰۲	۱۱۹	۱۹۱	۵۵۷	۳۸۷	۲۸۴
شی	**Worktop**	**Window**	**washbasin**	**Wall**	**trees**	**Tree**
تعداد تصاویر	۵۱۹	۱۳۹۳	۱۶۲	۴۴۶۲	۱۴۳	۱۱۲
شی	**desk lamp**	**Cushion**	**curtain**	**Building**	**bed crop**	**Bed**
تعداد تصاویر	۵۵۲	۴۹۳	۸۷۰	۷۹۵	۲۳۳	۵۰۷

۴ - ۴ - ۲ مجموعه‌داده UIUC Sport[۱]

این پایگاه داده شامل ۸ دسته از رخدادهای[۲] ورزشی است. این مجموعه داده در مجموع شامل ۱۵۷۹ تصویر در دسته‌های متفاوت می‌باشد. دسته‌های این مجموعه داده شامل، قایقرانی[۳]، بدمینتون[۴]، چوگان[۵]،

---

[۱] http://vision.stanford.edu/lijiali/event_dataset/event_dataset.rar

[۲] Event

[۳] Rowing

[۴] Badminton

[۵] Polo

تنیس روی میز،[1] اسنوبرد،[2] کروکت،[3] کشتیرانی[4] و صخره‌نوردی[5] می‌باشند. تصاویر این مجموعه داده به دو دسته آسان و متوسط بر اساس قضاوت موضوع توسط انسان، تقسیم می‌شوند. همچنین اطلاعات متنوع از پس‌زمینه اشیاء در هر تصویر موجود است. در این مجموعه ۱۰۰ تصویر از هر دسته برای آموزش، ۵۰ تصویر برای اعتبارسنجی و مابقی برای مجموعه آزمون استفاده گردید. لازم به ذکر است که ۱۴ شی با بالاترین تکرار در این مجموعه داده در نظر گرفته می‌شود.

۴ - ۴ - ۳ مجموعه داده Scence_15[6]

این مجموعه داده شامل ۱۵ کلاس از صحنه‌های مختلف است که ۸ دسته آن توسط مرجع [۱۰۹]، ۵ دسته توسط مرجع [۷۱] و دو دسته دیگر توسط مرجع [۸] جمع‌آوری شده است. ۱۵ دسته ذکر شده شامل صحنه‌های مختلف، اداری، آشپزخانه، اتاق خواب ، فروشگاه، صنعتی، ساختمان بلند، داخل شهر، خیابان‌ها، بزرگراه‌ها، سواحل، کوه، جنگل و حومه می‌باشد. اندازه تصویر در این پایگاه داده ۳۰۰*۲۵۰ است که در هر دسته ۲۱۰ تا ۴۱۰ تصویر وجود دارد. این مجموعه شامل بازه وسیعی از صحنه‌ها داخلی و خروجی از محیط است. در این مجموعه داده ۱۰۰ تصویر برای آموزش، ۵۰ تصویر برای اعتبارسنجی و مابقی برای آزمون در نظر گرفته می‌شود. برای این مجموعه داده ۲۷ شی با بالاترین تکرار در نظر گرفته شده است.

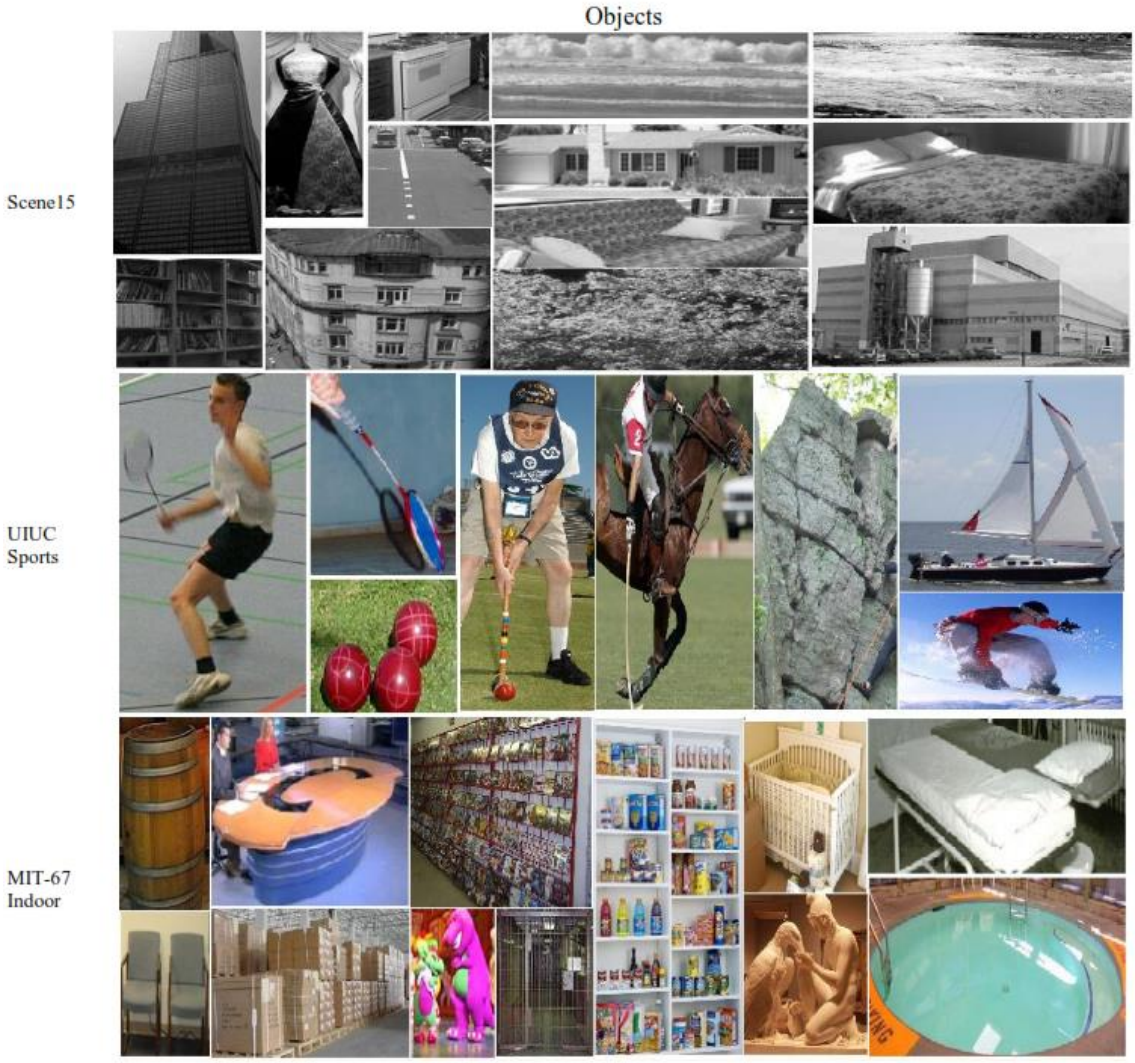
---

[1] Bocce
[2] Snowboarding
[3] Croquet
[4] Sailing
[5] Rock Climbing
[6] http://www-cvr.ai.uiuc.edu/ponce_grp/data/scene_categories/scene_categories.zip

۴ - ۴ - ۴ مجموعه داده MIT-67[1]

این مجموعه داده شامل ۶۷ دسته مختلف در محیط بسته می‌باشد. تعداد تصاویر در این مجموعه داده ۱۵۶۲۰ تصویر می‌باشد. حدوداً در هر دسته ۱۰۰ تصویر وجود دارد که ۷۰ تصویر برای مجموعه آموزش، ۱۰ تصویر برای اعتبارسنجی و مابقی برای آزمون در نظر گرفته می‌شود. برای این مجموعه داده ۱۰۳ شی با بالاترین تکرار در نظر گرفته شده است.

در شکل ۴-۲ نمونه‌هایی از مجموعه داده‌های استفاده شده برای دسته‌بندی صحنه نشان داده شده است.



شکل ۴-۲ نمونه‌های از سه مجموعه داده Scene 15، UIUC Sports و MIT-67 Indoor

[1] http://groups.csail.mit.edu/vision/LabelMe/NewImages/indoorCVPR_09.tar

۴ - ۴ - ۵ معیار ارزیابی برای دسته‌بندی صحنه پیچیده

معیارهای دقت و یادآوری مهم‌ترین و رایج‌ترین پارامترها برای ارزیابی و مقایسه سامانه برچسب‌گذاری تصاویر هستند. اما معیار دیگری به نام میانگین دقت متوسط[۱] (mAP)، کیفیت رتبه‌بندی تصاویر توسط روش را ارزیابی می‌کند. معیار ارزیابی استفاده شده در این بخش میانگین دقت متوسط است [۲۸].

رتبه‌بندی تصاویر یک ویژگی مهم و ذاتی برای روش‌های بازیابی تصاویر است. میانگین دقت متوسط عملکرد بازیابی را ارزیابی می‌کند. اگر تصاویر مربوط در رتبه‌های اول بازیابی شوند، این معیار به یک نزدیک خواهد شد و هر چه تصاویر مربوط در رتبه‌های آخر قرار گیرند، این معیار کمتر خواهد بود.

برای محاسبه میانگین دقت متوسط، ابتدا دقت متوسط (AP) برای پرس و جوی q محاسبه می‌شود. دقت متوسط همانند مرجع [۱۱۰] با رابطه‌ی (۴-۴) محاسبه می‌شود. در این رابطه i رتبه در دنباله بازیابی شده، n تعداد تصاویر بازیابی شده، $P(i)$ دقت بازیابی در i تصویر اول (تعداد تصاویر مرتبط تا رتبه iام تقسیم بر i) و $rel(i)$ یک تابع است. مقدار این تابع، با قرارگیری تصویر مرتبط با پرس‌وجو در رتبه iام، برابر یک و در غیر این صورت برابر صفر است. پارامتر $|releventimages|$ تعداد تصاویر استفاده شده در پایگاه داده را نشان می‌دهد.

$$AP(q) = \frac{\sum_{i=1}^{n}(P(i) \times rel(i))}{|releventimages|} \qquad (۴-۴)$$

میانگین دقت متوسط، با رابطه‌ی (۴-۵) محاسبه می‌شود. در این رابطه، Nq تعداد پرس‌وجوها را نشان می‌دهد

$$mAP = \frac{\sum_{i=1}^{n} AP(q)}{N_q} \qquad (۵-۴)$$

[۱] mean Average Precision (mAP)

در مرحله محاسبه‌ی نواحی مربوط به یک شی، از میزان هم‌پوشانی نواحی استفاده گردید. میزان مقبولیت نواحی با استفاده از رابطه (۴-۶) محاسبه می‌شود [۱۷].

$$R = \frac{\left(B_p \cap B_{g_t}\right)}{\left(B_p \cup B_{g_t}\right)} \qquad (۴-۶)$$

در رابطه‌ی (۴-۶)، B_p نواحی پیش‌بینی شده و B_{g_t} مرتبط با نواحی مطلوب است. با افزایش میزان R، دقت سیستم بالا می‌رود. در این رساله، میزان همپوشانی بیش از ۷۰٪ مطلوب در نظر گرفته می‌شود. برای محاسبه عملکرد روش دسته‌بندی صحنه از معیار صحت[۱] استفاده می‌شود. این معیار نسبت تعداد تصاویر درست تشخیص داده شده به تعداد کل تصاویر می‌باشد.

۴ - ۵ آزمایش‌ها و نتایج روش ترکیبی بر پایه Boosting و نواحی کاندیدا

همانطور که در فصل قبل بیان شد، برای تصاویر با پیچیدگی زیاد دو روش معرفی شده است. در روش اول از روش‌های یادگیری سنتی استفاده شد. در این روش از روش Boosting و نواحی کاندیدا برای دسته‌بندی صحنه استفاده شده است. در ادامه مراحل پیاده‌سازی و نتایج آزمایشات توسط این روش بیان می‌شود. لازم به ذکر است روش پیشنهادی بر روی مجموعه داده SUN مورد ارزیابی قرار گرفت. در آزمایش‌های انجام شده توسط روش پیشنهادی، از هسته گوسین $K(x,y) = e^{-\|x-y\|^2/2\sigma^2}$ برای SVM استفاده و تعداد نواحی کاندیدای استخراج شده توسط روش MCG برای هر تصویر ثابت و ۲۰ در نظر گرفته می‌شود.

[۱] Accuracy

۴ - ۵ - ۱ نتایج پیاده‌سازی روش Boosting و نواحی کاندیدا

روش پیشنهادی دارای چندین مرحله می‌باشد که دقت هر مرحله در دسته‌بندی صحنه تصویر بسیار تاثیرگذار می‌باشد. در همین راستا مراحل مختلف به صورت مجزا مورد تحلیل قرار گرفته می‌شود. در گام نخست دقت روش پیشنهادی برای برچسب‌گذاری ۳۶ شی با دیدگاه استفاده از روش نواحی کاندیدا مورد تحلیل قرار می‌گیرد. نتایج روش پیشنهادی در جدول ۴-۶ نشان داده شده است. همانطور که از نتایج به دست آمده مشخص می‌باشد روش پیشنهادی برای تشخیص اشیاء عملکرد بهتری نسبت به روش‌های قبل از خود را نشان داده است. لازم به ذکر است که پیاده‌سازی‌های انجام شده برای روش‌های مختلف از نویسندگان مربوطه دریافت شد و بر روی سخت افزار یکسان تست شده است.

جدول ۴-۶ مقایسه روش‌های تشخیص اشیاء بر پایه معیار mAP

روش / معیار ارزیابی	DPM v5 [112]	Selective Search[۱۹]	Regionlets [111]	روش پیشنهادی Boostingبرپایه
mAP	٪۳۰,۱۱	٪۳۳,۲۴	٪۳۸,۴۴	٪۴۲,۱۰

در ادامه برای دسته‌بندی صحنه از SVM استفاده می‌شود. این روش برپایه برچسب و ویژگی‌های استخراج شده از نواحی مرتبط به اشیاء می‌باشد. نتایج این مرحله در جدول ۴-۷ نشان داده شده است. همانطور که در نتایج جدول ۴-۷ مشاهده می‌شود روش پیشنهادی عملکرد بسیار مناسبی در مقایسه با روش‌های دیگر از خود نشان داده است. لازم به ذکر است که پیاده‌سازی‌های انجام شده برای روش‌های دیگران از نویسندگان مقاله دریافت شده و بر روی سخت افزار یکسان تست شده است.

جدول ۴-۷ مقایسه روش‌های دسته‌بندی صحنه تصویر بر پایه معیار صحت

روش / معیار ارزیابی	GIST-color [109]	RBoW [۱۱۳]	Object Bank [68]	روش پیشنهادی
صحت	٪۸۴,۲۳	٪۶۶,۷	٪۸۸,۹	**٪۹۳,۳**

در این بخش یک، روش نوین برای دسته‌بندی صحنه تصویر پیشنهاد شده است. روش پیشنهادی برپایه چند گام اساسی می‌باشد. در گام نخست برای محاسبه دقیق‌تر محدوده اشیاء و بجای جستجو کل تصویر از روش MCG برای پیشنهادی نواحی کاندیدا استفاده شده است. این روش سرعت پردازش در مراحل

بعدی را بسیار بهبود می‌دهد. در گام بعدی از روش Boosting بر پایه ویژگی‌های استخراج شده از پیچش و SVD برای تشخیص برچسب اشیاء استفاده شده است. در نهایت از روش SVM برای دسته-بندی صحنه تصاویر استخراج شده از مجموعه داده SUN استفاده می‌شود. نتایج نشان دهنده بهبود عملکرد روش پیشنهادی برای دسته‌بندی صحنه در این مجموعه داده می‌باشد. مشکل اساسی این روش زمان اجرای این روش می‌باشد. زمان آزمون و آزمایش آن بسیار زیاد می‌باشد و در واقع نمی‌توان از این روش برای کاربردهای واقعی استفاده نمود. به طور متوسط زمان مرحله آزمایش برای هر تصویر ۴ ثانیه می‌باشد.

۴ - ۶ آزمایش روش پیشنهادی بر پایه مدل موضوعی

در این بخش به معرفی روش پیشنهادی بر پایه مدل موضوعی و تحلیل بخش‌های مختلف این روش پرداخته می‌شود. لازم به ذکر است که این روش شامل چند بخش است که در گام نخست آزمایشات مرتبط به تشخیص و برچسب‌زنی اشیاء معرفی می‌شود. در گام بعدی، روش پیشنهادی برای دسته‌بندی صحنه بر پایه مدل موضوعی بیان می‌گردد.

در آزمایشات انجام شده توسط روش پیشنهادی، از هسته گوسین $K(x,y) = e^{-\|x-y\|^2/2\sigma^2}$ برای SVM فازی دوکلاسه و SVR استفاده شده است.

۴ - ۶ - ۱ نتایج آزمایش‌ها در تشخیص و برچسب‌گذاری اشیاء

در ابتدا روش‌های دیگران برپایه یادگیری عمیق در تشخیص اشیا معرفی می‌شود که پیاده‌سازی آن‌ها در یک سخت‌افزار مشابه اجرا شده است. در آزمایش‌های انجام شده برای هر شبکه عمیق از وزن‌های اولیه پیش‌آموزش داده شده بر روی مجموعه داده ImageNet استفاده شده است. در روش Faster R-CNNT از شبکه‌های عمیق با معماری مختلف استفاده شد. از مهم‌ترین معمارهای‌های عمیق استفاده شده می‌توان به VGG، ZF و ResNet اشاره نمود. در این رساله برای مقایسه بهتر از چهار روش معروف Fast RCNN [۱۶] ، Faster R-CNN [19] [۱۷]، SPPNet [۱۵]و R-FCN [۵۲] استفاده گردید. در

این بخش ابتدا نتایج به دست آمده از مجموعه داده SUN تحلیل می‌شود و در ادامه نتایج سه مجموعه داده دیگر معرفی می‌شوند. نتایج حاصل از آزمایش‌ها بر روی ۳۶ شی مجموعه داده SUN در جدول ۸-۴ برای هر شی مشاهده می‌شود. همانطور که در این جدول نشان داده شده است، روش R-FCN نسبت به سایر روش‌ها نتایج بهتری دارد. همچنین روش پیشنهادی بهترین نتایج را از خود نشان می‌دهد.

جدول ۸-۴ مقایسه روش‌های مختلف بر حسب mAP برای هر شی مجموعه داده SUN

روش / شی	Fast RCNN [۱۶]	Faster RCNN [۱۷]	SPPNet [۱۵]	R-FCN [۵۲]	Proposed Method
Bed	٪۷۸/۸۸	٪۷۹/۳۱	٪۷۸/۵۸	٪۷۶/۵۴	٪۷۷/۴۴
bed crop	٪۴۲/۵۸	٪۴۶/۱۶	٪۴۰/۴۳	٪۵۴/۶۵	٪۵۲/۸۳
Building	٪۴۵/۷۰	٪۴۸/۲۶	٪۴۵/۵۶	٪۴۸/۳۹	٪۴۹/۰۰
Buildings	٪۱۷/۴۲	٪۲۰/۹۱	٪۱۹/۱۸	٪۲۲/۹۴	٪۲۵/۱۴
Cabinet	٪۲۹/۴۶	٪۳۵/۱۹	٪۳۲/۶۷	٪۳۳/۸۵	٪۳۳/۱۵
Ceiling	٪۷۶/۴۷	٪۷۶/۲۵	٪۷۴/۹۱	٪۷۷/۸۵	٪۷۷/۹۶
Chair	٪۴۱/۱۸	٪۴۴/۵۰	٪۳۸/۵۲	٪۴۷/۱۸	٪۴۹/۰۱
chair occluded	٪۱۸/۸۴	٪۲۰/۹۵	٪۱۷/۸۶	٪۳۰/۲۳	٪۳۲/۰۱
Countertop	٪۴۲/۸۷	٪۴۶/۷۵	٪۴۰/۰۴	٪۴۶/۵۷	٪۵۲/۱۸
Curtain	٪۴۹/۴۷	٪۵۲/۲۰	٪۵۰/۴۵	٪۵۵/۶۴	٪۵۵/۲۸
Cushion	٪۳۷/۲۷	٪۳۵/۲۲	٪۳۳/۹۹	٪۵۱/۸۲	٪۵۲/۵۳
desk lamp	٪۵۶/۸۶	٪۵۹/۰۴	٪۵۵/۱۶	٪۷۰/۶۲	٪۷۰/۸۰
Faucet	٪۶/۰۷	٪۹/۱۲	٪۵/۱۰	٪۲۴/۶۶	٪۲۸/۷۹
Floor	٪۷۲/۰۶	٪۷۱/۸۸	٪۷۳/۵۷	٪۷۵/۵۸	٪۷۷/۲۱
Flowers	٪۲۵/۱۷	٪۲۷/۴۱	٪۲۴/۸۳	٪۳۴/۲۳	٪۴۲/۳۷
Microwave	٪۲۵/۸۴	٪۳۳/۶۸	٪۲۰/۹۰	٪۴۵/۸۶	٪۴۸/۰۷
Mirror	٪۳۲/۶۷	٪۳۵/۱۵	٪۲۹/۹۷	٪۳۰/۶۹	٪۳۵/۱۰
night table	٪۶۸/۵۵	٪۶۸/۰۱	٪۶۷/۱۲	٪۶۷/۸۸	٪۷۴/۰۴
night table occluded	٪۲۸/۵۷	٪۳۸/۱۹	٪۳۲/۶۴	٪۵۲/۳۸	٪۵۳/۸۰
Painting	٪۴۴/۷۵	٪۴۴/۹۲	٪۴۲/۹۵	٪۴۸/۳۴	٪۴۸/۷۸
Pillow	٪۳۰/۲۷	٪۳۲/۸۱	٪۳۲/۶۳	٪۴۴/۱۳	٪۴۷/۰۲
pillow occluded	٪۱۳/۹۰	٪۶/۸۲	٪۱۲/۰۱	٪۱۷/۳۰	٪۲۱/۳۷
Plant	٪۲۴/۰۳	٪۲۴/۵۶	٪۲۴/۱۱	٪۳۰/۲۲	٪۳۴/۸۹
Sink	٪۲۵/۰۴	٪۲۱/۸۲	٪۲۳/۲۵	٪۲۵/۶۴	٪۳۱/۸۸
Sky	٪۸۹/۸۶	٪۸۹/۵۹	٪۸۹/۶۲	٪۹۰/۱۴	٪۹۰/۱۳
Skyscraper	٪۶۰/۹۸	٪۶۱/۵۱	٪۵۷/۴۱	٪۶۵/۷۲	٪۶۷/۱۲

Stove	٪۳۰/۱۷	٪۳۰/۱۳	٪۳۳/۷۹	٪۲۹/۱۳	٪۲۷/۴۱
Table	٪۲۷/۶۲	٪۲۹/۱۱	٪۲۸/۳۴	٪۲۵/۶۹	٪۲۶/۰۷
Toilet	٪۵۶/۱۴	٪۶۹/۲۲	٪۵۳/۵۱	٪۶۶/۱۷	٪۶۶/۵۹
Towel	٪۸/۲۴	٪۱۵/۸۶	٪۱۰/۵۳	٪۲۱/۲۹	٪۲۰/۲۵
Tree	٪۲۹/۴۵	٪۳۱/۵۲	٪۳۰/۲۲	٪۳۴/۵۲	٪۳۵/۰۴
Trees	٪۴۳/۲۱	٪۴۴/۸۶	٪۴۰/۷۹	٪۴۶/۶۳	٪۴۹/۳۸
Wall	٪۵۹/۴۷	٪۶۱/۶۳	٪۵۹/۸۰	٪۶۲/۶۱	٪۶۳/۶۹
Washbasin	٪۲۳/۱۷	٪۳۶/۶۲	٪۲۱/۲۹	٪۴۰/۰۷	٪۳۳/۶۴
Window	٪۳۴/۰۱	٪۳۶/۶۱	٪۳۲/۲۴	٪۴۵/۳۶	٪۴۸/۰۹
Worktop	٪۲۸/۲۲	٪۳۳/۸۲	٪۲۳/۷۷	٪۳۹/۱۹	٪۴۳/۶۸
متوسط	٪۳۹/۵۷	٪۴۲/۲۲	٪۳۸/۸۳	٪۴۶/۶۶	٪۴۸/۳۸

روش پیشنهادی، بهبود یافته روش R-FCN می‌باشد که از SVM فازی دو کلاسه با SVR بجای رای‌گیری استفاده شده است. همچنین از تابع زیان جدید استفاده شده است. در جدول ۴-۹ مقایسه‌ای بین روش‌های معرفی شده با معماری‌های متفاوت انجام شده است. همانطور که در این جدول مشاهده می‌شود، روش پیشنهادی با معماری ResNet در مقایسه با روش‌های دیگر بهترین نتایج را از خود نشان می‌دهد.

جدول ۴-۹ مقایسه روش های مختلف با معماری‌های مختلف در مجموعه داده SUN در تشخیص اشیاء صحنه با معیار mAP

معماری / روش	ResNet	ZF	VGG
Fast RCNN [۱۶]	٪۳۸/۲۱	٪۳۲/۱۷	٪۳۹/۵۷
Faster RCNN [۱۷]	٪۴۰/۲۳	٪۳۹/۰۹	٪۴۲/۲۲
SPPNet [۱۵]	٪۳۷/۹۸	٪۳۸/۸۳	٪۳۶/۳۳
R-FCN [۵۲]	٪۴۶/۶۶	٪۴۱/۰۷	٪۴۲/۹۹
Proposed Method	**٪۴۸/۳۸**	٪۴۴/۶۷	٪۴۵/۰۸

مقایسه دیگری که در این بخش انجام شد، بر اساس اندازه لایه‌های ResNet و تعداد پیچش است. معمولاً این شبکه با تعداد لایه‌های ۵۰، ۱۰۱ و ۱۵۲ در شناسایی شی استفاده می‌شود و وزن‌های اولیه برای این شبکه با توجه به مجموعه ImageNet موجود می‌باشد. در همین راستا، این تعداد لایه برای آزمایش انتخاب شده است. همانطور که در جدول ۴-۱۰ نشان داده شد روش پیشنهادی با معماری ResNet با ۱۰۱ لایه بهترین نتایج را ارائه می دهد.

جدول ۴-۱۰ مقایسه معماری ResNet با تعداد لایه متفاوت در مجموعه داده SUN در تشخیص اشیاء صحنه با معیار mAP

تعداد لایه / روش	ResNet-50	ResNet-101	ResNet-152
R-FCN [۵۲]	٪۴۳/۲۱	٪۴۶/۶۶	٪۴۶/۰۵
Proposed Method	٪۴۵/۹۳	**٪۴۸/۳۸**	٪۴۷/۶۷

همانطور که اشاره شد، استفاده از تابع زیان مناسب در افزایش دقت عملکرد شبکه عمیق بسیار موثر است. در این راستا در جدول ۴-۱۱ مقایسه‌ای بین روش R-FCN با تابع زیان‌های متفاوت انجام شده است. نتایج جدول ۴-۱۱ نشان می‌دهد که روش R-FCN با تابع زیان اختلاف کوشی-شوارتز نسبت به آنتروپی متقابل، عملکرد بهتری از لحاظ دقت داشته است. همچنین در این جدول مشاهده می‌شود که متوسط زمان اجرا برای هر تصویر در تابع زیان اختلاف کوشی-شوارتز بهتر از زمان R-FCN با تابع زیان آنتروپی متقابل است. معماری استفاده شده برای هر دو شبکه عمیق، ResNet با ۱۰۱ لایه است.

جدول ۴-۱۱ مقایسه تابع زیان در شبکه R-FCN در مجموعه داده SUN در تشخیص اشیاء صحنه با معیار mAP

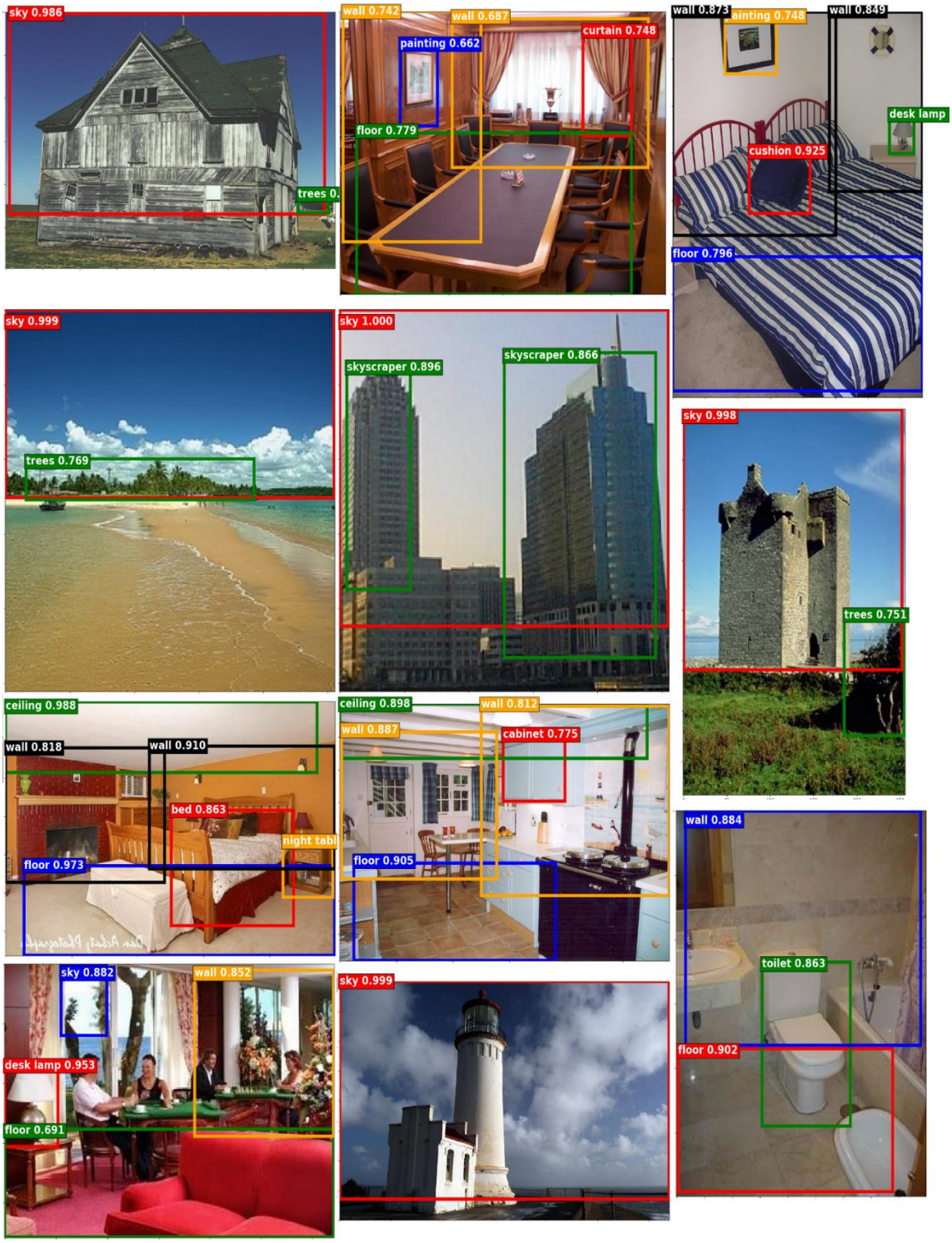
روش	R-FCN [۵۲]	R-FCN With Cauchy-Schwarz Divergence Loss
mAP	٪۴۶/۶۶	**٪۴۷/۰۴**
زمان (ثانیه)	۰,۱۷	**۰,۱۱**

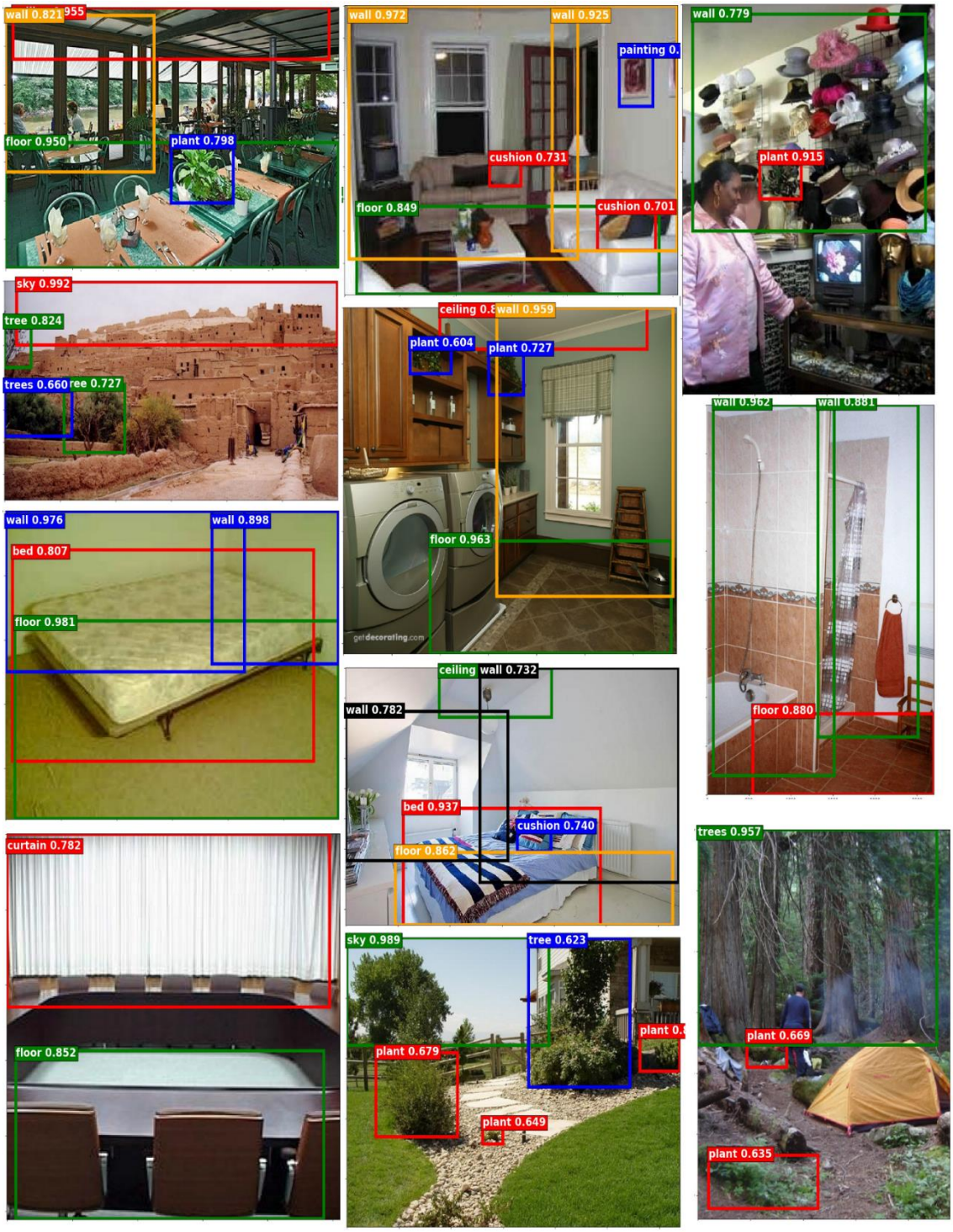
زمان مرحله آزمون، یک معیار مهم و اساسی در بررسی روش‌های آشکارسازی و برچسب‌زنی تصاویر است. در همین راستا، جدول ۴-۱۲ مقایسه‌ای بین روش‌های مطرح شده و روش پیشنهادی بر اساس متوسط زمان برای هر تصویر را نشان می‌دهد. همانطور که در جدول ۴-۱۲ نشان داده می‌شود، روش پیشنهادی از نظر زمان اجرا، عملکرد بهتری را داشته است. نتایج جدول ۴-۸، جدول ۴-۱۱ و جدول ۴-۱۲ ، مبین این نکته می‌باشد که متوسط زمان برای هر تصویر در روش پیشنهادی نسبت به R-FCN با تابع زیان اختلاف کوشی-شوارتز افزایش می‌یابد ولی همچنان از روش R-FCN با تابع زیان آنتزوپی متقابل بهتر می‌باشد و عملکرد بهتری از لحاظ دقت نسبت به بقیه روش‌ها دارد.

جدول ۴-۱۲ مقایسه زمان اجرای تصویر آزمون در مجموعه داده SUN

روش	Fast RCNN [۱۶]	Faster RCNN [۱۷]	SPPNet [۱۵]	R-FCN [۵۲]	Proposed Method
زمان (ثانیه)	۰/۴۹	۰/۴۳	۰/۳۸	/۱۷.	**۰/۱۳**

شکل ۴-۳ نمونه‌های از نتایج بدست آمده از R-FCN بهبود یافته شده را نشان می‌دهد. همانطور که در شکل مشاهده می‌شود، نتایج حاصل نشان‌دهنده عملکرد مناسب روش پیشنهادی در آشکارسازی و برچسب‌زنی اشیاء است.


sky 0.986
trees 0.
wall 0.742
wall 0.687
curtain 0.748
painting 0.662
floor 0.779
wall 0.873
wall 0.849
desk lamp
cushion 0.925
floor 0.796
sky 0.999
trees 0.769
sky 1.000
skyscraper 0.896
skyscraper 0.866
sky 0.998
trees 0.751
ceiling 0.988
wall 0.818
wall 0.910
bed 0.863
night tabl
floor 0.973
ceiling 0.898
wall 0.812
wall 0.887
cabinet 0.775
floor 0.905
wall 0.884
toilet 0.863
floor 0.902
sky 0.882
wall 0.852
desk lamp 0.953
floor 0.691
sky 0.999



شکل ۴-۳ نمونه‌های از خروجی بدست آمده در مجموعه داده SUN برای تشخیص و برچسب زنی اشیاء موجود در تصویر

در ادامه نتایج بدست آمده بر روی سه مجموعه داده MIT-67، UIUC Sports و Scene-15 تحلیل می‌شود. همانطور که ذکر شده تعداد اشیاء موجود در این مجموعه داده به ترتیب ۱۰۳، ۲۷ و ۱۴ می‌باشد. نتایج به دست آمده در این سه مجموعه داده مشابه مجموعه داده قبل می‌باشد.

در ابتدا مقایسه‌ای بین معماری‌های مختلف شبکه‌های پیچش برای برچسب زنی اشیاء انجام می‌شود. در جدول ۴-۱۳نتایج بدست آمده از روش پیشنهادی برای تشخیص و برچسب‌زنی اشیاء در سه مجموعه داده با معماری‌های متفاوت را نشان می‌دهد. معماری تمام پیچش ResNet بهترین عملکرد را از خود نشان داده است.

جدول ۴-۱۳ مقایسه روش پیشنهادی با معماری‌های مختلف در مجموعه داده‌های Scene15، UIUC Sports و MIT-67 در تشخیص اشیاء صحنه با معیار mAP

معماری / مجموعه داده	ResNet	ZF	VGG
Scene15	**٪۷۴,۹**	٪۵۹,۱	٪۶۱,۷۷
UIUC Sports	**٪۸۹,۴۵**	٪۶۸,۱	٪۷۷,۲۱
MIT-67 Indoor	**٪۵۹,۳۲**	٪۴۴,۵	٪۵۱,۳۳

در معماری ResNet مقایسه بین تعداد لایه‌ها در جدول ۴-۱۴ نشان داده شده است. این معماری با ۱۰۱ لایه بهترین عملکرد را از خود نشان داده است.

جدول ۴-۱۴مقایسه روش پیشنهادی با تعداد لایه‌های متفاوت معماری ResNet در مجموعه داده‌های Scene15، UIUC Sports و MIT-67 در تشخیص اشیاء صحنه با معیار mAP

تعداد لایه / مجموعه داده	ResNet-50	ResNet-101	ResNet-152
Scene15	٪۷۰,۲۳	**٪۷۴,۹**	٪۷۳,۲۱
UIUC Sports	٪۸۵	**٪۸۹,۴۵**	٪۸۶,۷
MIT-67 Indoor	٪۵۳,۷	**٪۵۹,۳۲**	٪۵۷,۱۲

جدول ۴-۱۵ نتایج به دست آمده از با تابع زیان متفاوت در R-FCN معرفی شده در مرجع [۵۲] را نشان می‌دهد. همانطور که مشاهده می‌شود در هر سه مجموعه داده تابع زیان کوشی-شوارتز بهترین عملکرد را داشته است.

جدول ۴-۱۵ مقایسه تابع زیان در روش R-FCN در مجموعه داده‌های Scene15، UIUC Sportو MIT-67 در تشخیص اشیاء صحنه با معیار mAP

روش‌ها / مجموعه داده	R-FCN With cross-entropy loss function	R-FCN with Cauchy-Schwartz loss function
Scene15	٪۶۳,۲۳	**٪۶۶,۷**
UIUC Sports	٪۷۷,۱۲	**٪۸۱,۱۵**
MIT-67 Indoor	٪۴۱,۷۰	**٪۴۸,۱۱**

جدول ۴-۱۶ مقایسه بین روش پیشنهادی بر اساس SVM فازی و SVR برای تشخیص اشیاء با روش‌های دیگری یادگیری عمیق را نشان می‌دهد که نشان‌دهنده بهتر بودن روش پیشنهادی نسبت به روش‌های دیگر می‌باشد.

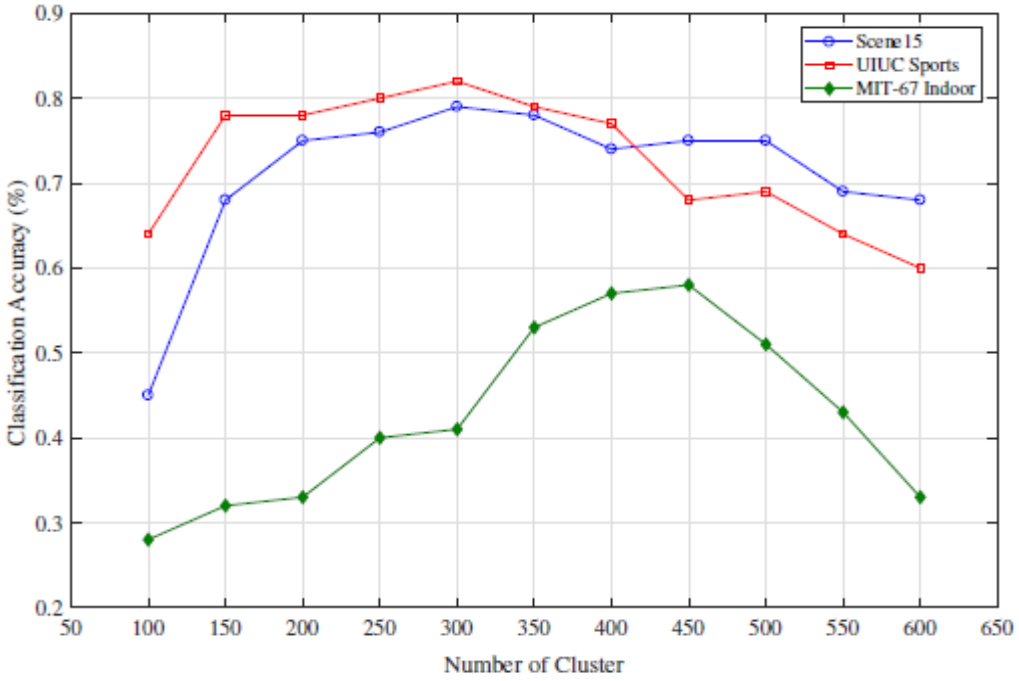
جدول ۴-۱۶ مقایسه روش‌های یادگیری عمیق برای تشخیص و برچسب‌زنی اشیاء در سه مجموعه داده Scene15، UIUC Sports و MIT-67 با معیار mAP

روش‌ها / مجموعه داده	Proposed Method	R-FCN	Fast RCNN	Faster R-CNN	SPPNet
Scene15	**٪۷۴,۹**	٪۶۳,۲۳	٪۵۹,۱۲	٪۶۱,۴۰	٪۵۸,۹
UIUC Sports	**٪۸۹,۴۵**	٪۷۷,۱۲	٪۷۱,۸۷	٪۷۲	٪۷۸,۸۶
MIT-67 Indoor	**٪۵۹,۱۳**	٪۴۱,۷۰	٪۳۹,۲۰	٪۴۴,۲۳	٪۴۵,۴

۴ - ۶ - ۲ استخراج مجموعه کلمات بصری و ویژگی‌های موضوعی

همانطور که در روش پیشنهادی بیان شد، بعد از استخراج نواحی و برچسب مرتبط به اشیاء باید کلمات بصری و ویژگی‌های موضوعی استخراج شود. در استخراج کلمات بصری تعداد خوشه‌ها برای ساخت کتابچه کد بسیار مهم می‌باشد. در همین راستا در این بخش با توجه به مجموعه اعتبارسنجی تعداد خوشه‌ها محاسبه شده است. در شکل ۴-۴ تعداد خوشه‌های استخراجی برای هر مجموعه داده نشان داده شده است. با توجه به این شکل تعداد خوشه برای سه مجموعه داده Scene-15، UIUC Sport و MIT-67 به ترتیب ۳۰۰، ۳۰۰ و ۴۵۰ می‌باشد. برچسب‌های استفاده شده برای مجموعه کلمات بصری نیز برابر با این تعداد در هر سه مجموعه داده در نظر گرفته می‌شود. در ادامه آزمایش‌های انجام شده، تعداد خوشه‌های برای هر سه مجموعه داده ثابت و برابر با مقادیر محاسبه شده فرض می‌شود. برای به

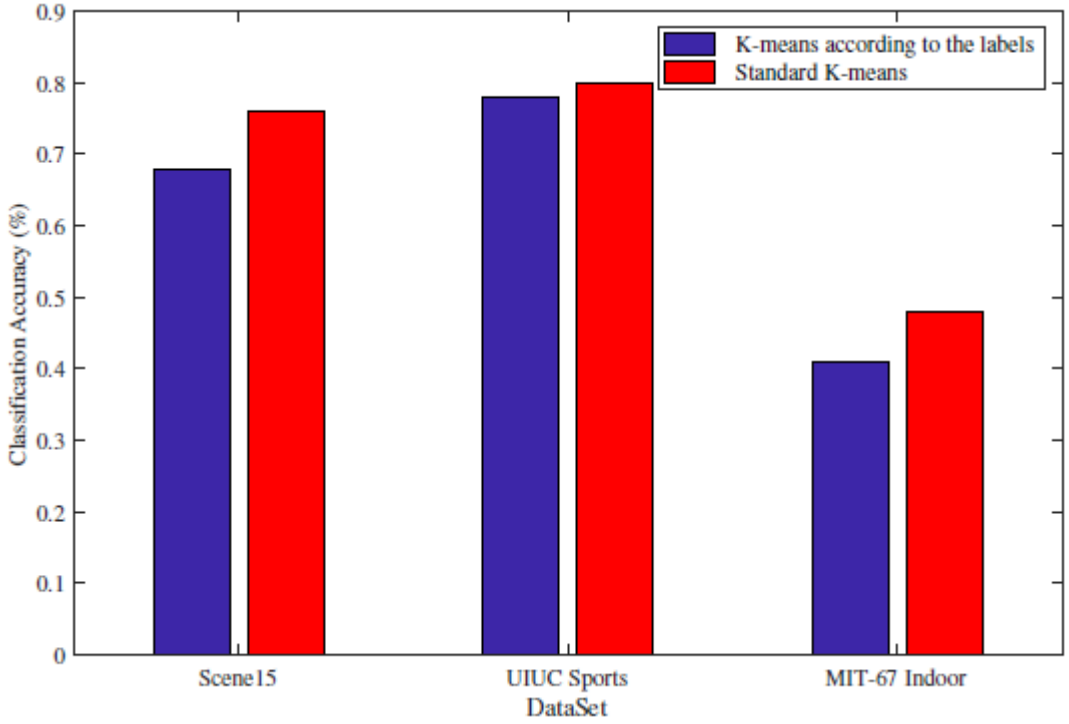
دست آوردن میزان دقت از روش تخمین‌گر توزیع‌شده خودرگرسیو عصبی اسناد استفاده شد و تعداد ویژگی‌های موضوعی ۱۰۰ در نظر گرفته شده است. در این روش از ویژگی‌های استخراج شده از کلمات بصری توسط SIFT و برچسب نواحی استفاده می‌شود. همچنین برای کاهش کلمات بصری از روش PCA استفاده و برای محاسبه تعداد ویژگی از رابطه (۴-۷) استفاده می‌شود.



شکل ۴-۴ تعداد خوشه‌های استخراجی برای سه مجموعه داده Scene15، UIUC Sports و MIT-67

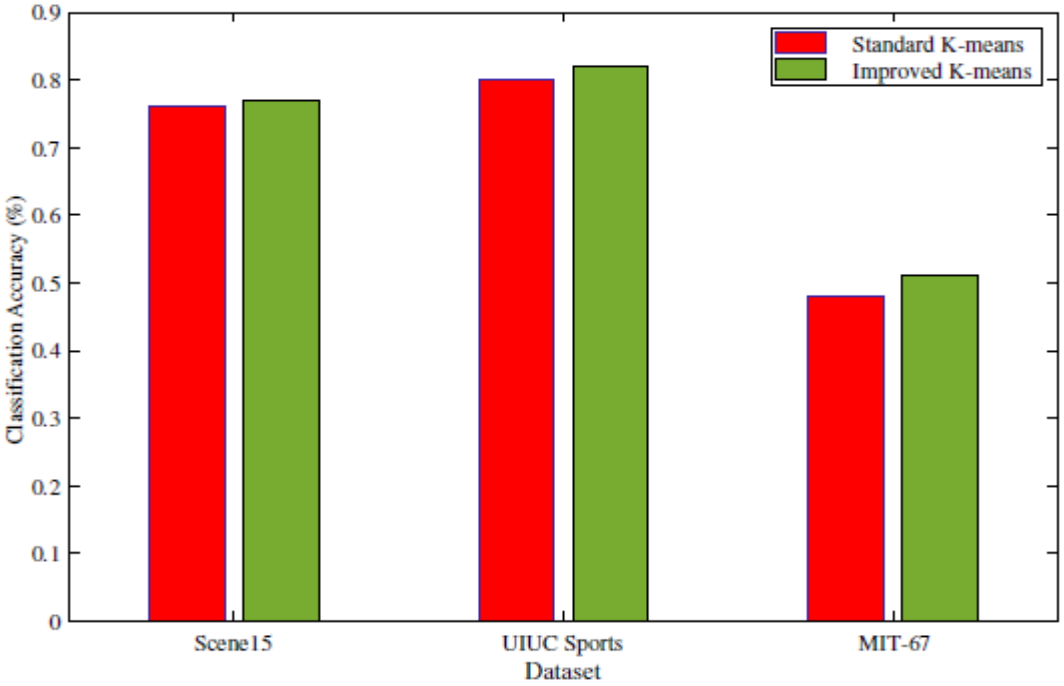
$$FeaturePCA = \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{j=1}^{n} \lambda_j} \geq 0.98 \qquad (۴-۷)$$

در این رابطه λ نشان دهنده مقادیر ویژه، k تعداد ویژگی‌های انتخابی و n تعداد کل ویژگی‌ها می‌باشد. همانطور که در روش پیشنهادی بیان شد، برای خوشه‌بندی می‌توان دو دیدگاه با توجه به برچسب نواحی استخراج در نظر گرفت. در دیدگاه اول خوشه‌بندی براساس برچسب ناحیه انجام می‌شود. در واقع برای نواحی با برچسب یکسان، خوشه‌بندی جداگانه انجام می‌شود و در دیدگاه دوم خوشه‌بندی بدون برچسب در نظر گرفته می‌شود. یعنی تمامی نواحی دارای ارزش یکسان می‌باشد. شکل ۴-۵ مقایسه‌ای بین این دو دیدگاه را نشان می‌دهد. برای خوشه‌بندی از k-means استاندارد استفاده شده است. با توجه به نتیجه به دست آمده از این شکل دیدگاه دوم عملکرد بهتری دارد.



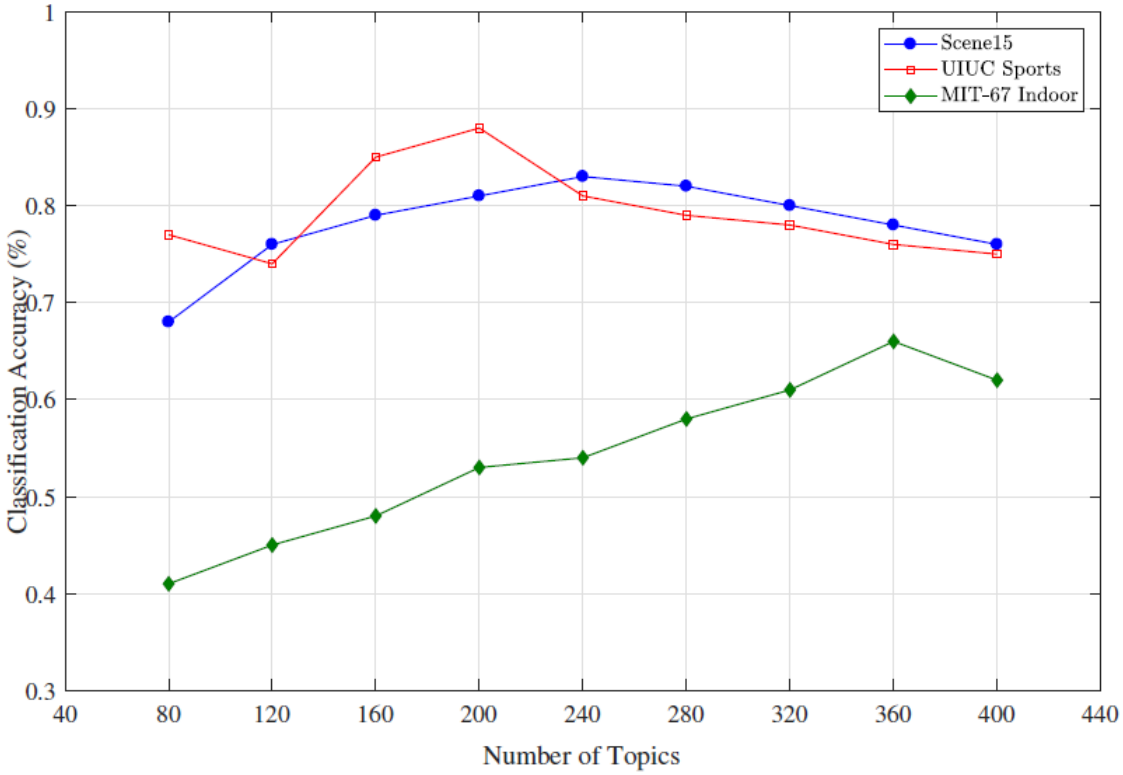
شکل ۴-۵ مقایسه خوشه‌بندی بر اساس برچسب ناحیه و بدون در نظر گرفتن برچسب ناحیه

در این رساله از یک k-mean توسعه یافته برای بالا بردن دقت روش پیشنهادی استفاده شده است. شکل ۴-۶ مقایسه‌ای بین دو روش خوشه بندی را نشان می‌دهد. همانطور که در این شکل نشان داده شده k-means توسعه یافته عملکرد بهتری نسبت به k-means استاندارد از خود نشان داده است.



شکل ۴-۶ مقایسه روش خوشه‌بندی در مجموعه کلمات بصری در سه مجموعه داده Scene15، UIUC Sports و MIT-67

روش تخمین‌گر توزیع‌شده خودرگرسیو عصبی اسناد به تعداد ویژگی‌های موضوعی وابسته می‌باشد. تاکنون تعداد این ویژگی‌ها ثابت و ۱۰۰ در نظر گرفته شده است. ولی در ادامه با توجه به مجموعه اعتبارسنجی تعداد بهینه این ویژگی‌ها برای هر مجموعه داده استخراج می‌شود. شکل ۴-۷ مقایسه تعداد ویژگی‌های موضوعی را برای سه مجموعه داده نشان می‌دهد. همانطور که مشاهده می‌شود روش تخمین‌گر توزیع‌شده خودرگرسیو عصبی اسناد برای سه مجموعه داده Scene-15، UIUC Sport و MIT-67 به ترتیب با ۲۴۰، ۲۰۰ و ۳۶۰ ویژگی موضوعی بهترین عملکرد را دارد.



شکل ۴-۷ تعداد ویژگی‌های موضوعی استخراج شده در روش تخمین‌گر توزیع‌شده خودرگرسیو عصبی اسناد

در مرحله آزمون از ساختار بهینه به دست آمده در این بخش برای روش تخمین‌گر توزیع‌شده خودرگرسیو عصبی اسناد استفاده می‌شود که این پارامترها در جدول ۴-۱۷نشان داده شده است.

جدول ۴-۱۷ مشخصات استفاده برای آموزش و آزمون روش تخمین‌گر توزیع‌شده خودرگرسیو عصبی اسناد

مجموعه داده / پارامترها	Scene-15	UIUC Sport	MIT-67
تعداد خوشه	۳۰۰	۳۰۰	۴۵۰
تعداد ویژگی های موضوعی	۲۴۰	۲۰۰	۳۶۰

۴ - ۶ - ۳ دسته‌بندی صحنه بر پایه روش مدل موضوعی

همانطور که در روش پیشنهادی بیان شد، برای دسته‌بندی صحنه از ۳ نوع ویژگی استفاده می‌شود. این ویژگی‌ها شامل ویژگی‌های استخراج شده از لایه آخر ResNet، ویژگی‌های موضوعی و برچسب مجموعه کلمات بصری استخراج شده از نواحی مرتبط به اشیاء می‌باشد. برای کاهش بعد ویژگی‌های استخراج شده از لایه پیچش از روش PCA استفاده می‌شود. تعداد این ویژگی‌ها از رابطه (۴-۷) محاسبه شده است. جدول ۴-۱۸ ویژگی‌های استفاده شده و تعداد هر نوع ویژگی را نشان می‌دهد. در پایان برای دسته‌بندی از یک دسته‌بند ساده بیزین استفاده می‌شود. برای استخراج ویژگی‌های پیچش از نواحی مرتبط به اشیاء استخراجی با توجه به محل قرار گرفتن استفاده شد.

جدول ۴-۱۸ ویژگی‌ها و تعداد آنها برای دسته‌بندی صحنه در روش پیشنهادی بر پایه مدل‌های موضوعی در سه مجموعه داده Scene-15، UIUC Sport و MIT-67

		ورودی			خروجی
		ResNet-Feature	Topic Feature	Label of BoVW	
مجموعه داده	Scene15	$PCA\begin{pmatrix}3\times3,Conv,512\\3\times3,Conv,512\end{pmatrix}=380$	240	300	15
	UIUC Sport	$PCA\begin{pmatrix}3\times3,Conv,512\\3\times3,Conv,512\end{pmatrix}=290$	200	300	8
	MIT-67 Indoor	$PCA\begin{pmatrix}3\times3,Conv,512\\3\times3,Conv,512\end{pmatrix}=500$	360	450	67

۴ - ۶ - ۴ مقایسه روش پیشنهادی بر پایه ویژگی‌های موضوعی با روش‌های دیگر

در این بخش به مقایسه روش پیشنهادی با روش‌های مطرح شده در سال‌های اخیر برای دسته‌بندی صحنه تصویر پرداخته می‌شود. لازم به ذکر است مجموعه داده و تعداد انتخاب شده در هر سه مجموعه داده برای همه‌ی روش‌ها یکسان می‌باشد و نتایج مرتبط با پژوهش دیگران از مقالات ایشان استخراج شد. روش‌هایی که برای مقایسه انتخاب گردید به گونه‌ای انتخاب شد که علاوه بر جدید بودن تمامی

روش‌های موجود در دسته‌بندی را شامل می‌شود. این روش‌ها شامل روش‌های سنتی، مدل‌های موضوعی و یادگیری عمیق بر اساس ویژگی‌های سطح بالا و سطح پایین می‌باشند.

جدول ۴-۱۹ نتایج به دست آمده از روش پیشنهادی و روش‌های موجود بر روی مجموعه داده UIUC Sports را نشان می‌دهد. همانطور که از نتایج این جدول مشخص می‌باشد روش پیشنهادی عملکرد بهتری نسبت به روش‌های موجود برخوردار است.

جدول ۴-۱۹ مقایسه روش پیشنهادی بر پایه مدل موضوعی در مجموعه داده UIUC Sport

UIUC 8-Sport Dataset	
روش‌ها	**صحت (%)**
GIST-color [۱۰۹]	۷۰,۷
MM-Scene [۷۵]	۷۱,۷
Graphical [۷۷]	۷۳,۴
Object Bank [۶۸] [68]	۷۶,۳
Object Attributes [۱۱۴]	۷۷,۹
CENTRIST [۱۱۵]	۷۸,۲
RSP [۱۱۶]	۷۹,۶
SPM [۸]	۸۱,۸
SPMSM [۱۱۷] [117]	۸۳,۰
Classemes [۱۱۸]	۸۴,۲
HIK [۱۱۹]	۸۴,۲
LScSPM [۱۲۰]	۸۵,۳
LPR-RBF [۱۲۱]	۸۶,۲
Hybrid Parts + GIST + SP [۱۲۲]	۸۷,۲
LCSR [۱۲۳]	۸۷,۲
VC + VQ [۱۲۴]	۸۸,۴
IFV [۱۲۵]	۹۰,۸
ISPR [۱۲۶]	۸۹,۵
DRCF [۱۲۷]	۹۸,۷
Wang with Xception [۱۲۸]	۹۸,۷
SMNN+Xception [۱۲۹]	۹۸,۹
Proposed Method	**۹۹,۲**

جدول ۴-۲۰ نتایج بدست آمده از روش پیشنهادی و روش‌های موجود بر روی مجموعه داده Scene-15 را نشان می‌دهد. همانطور که نتایج این جدول نشان می‌دهد روش پیشنهادی عملکرد بهتری نسبت به روش‌های موجود دارد.

جدول ۴-۲۰ مقایسه روش پیشنهادی بر پایه مدل موضوعی در مجموعه داده Scene-15

Scene-15 Dataset	
روش ها	**صحت (٪)**
GIST-color [۱۰۹]	۶۹,۵
RBoW [۱۱۳]	۷۸,۶
Classemes [۱۱۸]	۸۰,۶
Object Bank [۶۸]	۸۰,۹
SPM [۸]	۸۱,۴
SPMSM [۱۱۷]	۸۲,۳
LCSR [۱۲۳]	۸۲,۷
SP-pLSA [۱۳۰]	۸۳,۷
CENTRIST [۱۱۵]	۸۳,۹
HIK [۱۱۹]	۸۴,۱
OTC [۱۳۱]	۸۴,۴
ISPR [۱۲۶]	۸۵,۱
VC + VQ [۱۲۴]	۸۵,۴
LMLF [۶۷]	۸۵,۶
LPR-RBF [۱۲۱]	۸۵,۸
Hybrid Parts + GIST + SP [۱۲۲]	۸۶,۳
CENTRIST+LCC+Boosting [۱۳۲]	۸۷,۸
RSP [۱۱۶]	۸۸,۱
IFV [۱۲۵]	۸۹,۲
LScSPM [۱۲۰]	۸۹,۷
DRCF [۱۲۷]	۹۴,۵
SMNN-Xception [۱۲۹]	۹۶,۵
Wang With Xception [۱۲۸]	۹۷.۲
Proposed Method	**۹۸.۵**

جدول ۴-۲۱ نتایج به دست آمده از روش پیشنهادی و روش‌های موجود بر روی مجموعه داده MIT-67 را نشان می‌دهد. نتایج این جدول بیانگر بهبود روش پیشنهادی عملکرد نسبت به روش‌های دیگر می‌باشد. روش‌های که در سال‌های اخیر بر روی این مجموعه داده ارائه شده بر پایه روش‌های یادگیری عمیق بوده عملکرد بهتری نسبت به روش‌های سنتی و مدل‌های موضوعی داشته است.

جدول ۴-۲۱ مقایسه روش پیشنهادی بر پایه مدل موضوعی در مجموعه داده MIT-Sports

MIT-67 Indoor Dataset	
روش‌ها	**صحت (٪)**
ROI + GIST [۱۳۳]	۲۶,۱
MM-Scene [۷۵]	۲۸,۳
SPM [۸]	۳۴,۴
Object Bank [۶۸]	۳۷,۶
RBoW [۱۱۳]	۳۷,۹
Weakly Supervised DPM [۱۳۴]	۴۳,۱
SPMSM [۱۱۷]	۴۴,۰
LPR-LIN [۱۲۱]	۴۴,۸
BoP [۱۳۵]	۴۶,۱
Hybrid Parts + GIST + SP [۱۲۲]	۴۷,۲
OTC [۱۳۱]	۴۷,۳
Discriminative Patches [۱۳۶]	۴۹,۴
ISPR [۱۲۶]	۵۰,۱
D-Parts [۱۳۷]	۵۱,۴
VC + VQ [۱۲۴]	۵۲,۳
IFV [۱۲۵]	۶۰,۸
MLRep [۱۳۸]	۶۴,۰
CNN-MOP [۱۳۹]	۶۸,۹
CNNaug-SVM [۱۴۰]	۶۹,۰
DRCF [۱۲۷]	۷۱,۸
Deep Filter Banks [۱۴۱]	۷۴,۲
DVW [۱۴۲]	۷۷,۴
Yoo2015 [۱۴۳]	۸۰,۷۸

Wang with ResNet [۱۲۸]	۸۱,۶
DSPMK [۱۴۴]	۸۱,۸۷
Proposed Method	**۸۳,۲۱**

با توجه به نتایج بدست آمده در جدول ۴-۱۹ تا جدول ۴-۲۱ روش‌های ارائه شده در مراجع [۱۲۹] و [۱۲۸] عملکردهای بسیار مناسبی در سه مجموعه داده استفاده شده در این رساله از خود نشان داده‌اند. این دو مرجع جدیدترین روش‌های ارائه شده در شناسایی صحنه از لحاظ معنایی می‌باشند. در این مراجع ادعا شده برای بدست آوردن عملکرد مناسب در دسته‌بندی صحنه نمی‌توان تنها از اطلاعات مرتبط به تصویر استفاده کرد. در واقع بیان شد که استفاده تنها از الگوهای بصری در کل تصویر نمی‌تواند در دسته‌بندی صحنه موثر باشد دلیل آن را نیز چالش‌های موجود در تصویر مانند شباهت پس‌زمینه و شباهت‌های موجود در محل قرار گرفتن اشیاء ذکر کردند. در همین راستا از اطلاعات حاشیه‌نویسی انسان[۱] بدست آمده از موتور جستجو گوگل و اطلاعات مرتبط به کل تصویر استفاده شده است. در واقع برچسب‌های استخراجی از متن توضیحات تصاویر که توسط کاربران در صفحات وب موجود بوده به همراه کل تصاویر استفاده شده است. از دو معماری عمیق برای دسته‌بندی و ترکیب ویژگی‌های معنای سطح بالا استفاده شده است.

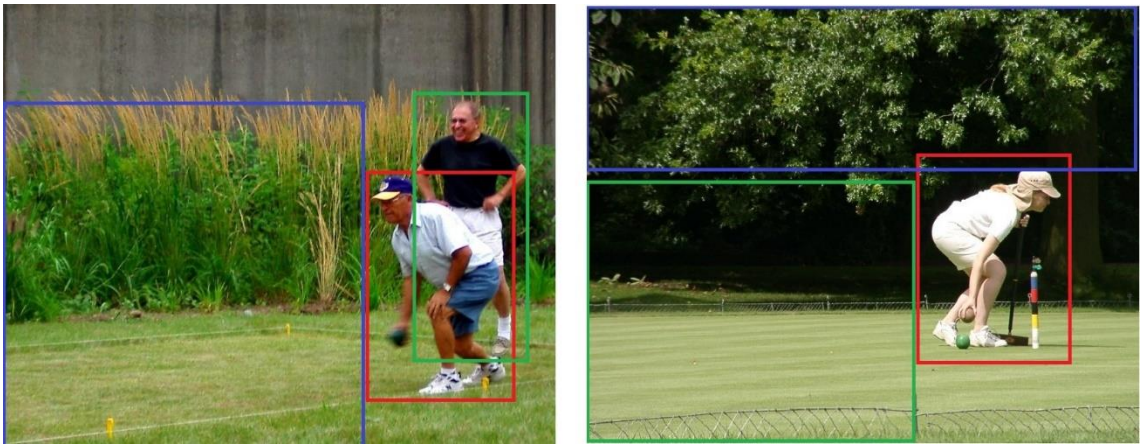
ولی در روش پیشنهادی ارائه شده در این رساله، نشان داده شد اگر بتوان اطلاعات مناسب از اشیاء موجود در تصویر و موضوعات نهفته در تصویر را استخراج نمود می‌توان از تصویر به تنهایی برای دسته‌بندی استفاده کرد. از طرفی استفاده از اطلاعات وب می‌تواند به دلیل خطای انسانی و متفاوت بودن شیوه نوشتاری برچسب‌ها در حاشیه‌نویسی انسانی این روش را با مشکل روبرو کند. نتایج بدست آمده از روش پیشنهادی رساله عملکرد بهتری نسبت به روش‌های دیگر داشته است.

[۱] Human-Annotated

مرجع [۱۲۸] برای نمونه شکل ۴-۸ را برای نمایش ناتوان بودن الگورهای بصری برای بدست آوردن صحت مناسب در دسته‌بندی صحنه ارائه کرده است. در این مرجع از ویژگی‌های استخراجی توسط یادگیری عمیق از کل تصویر استفاده شده است.

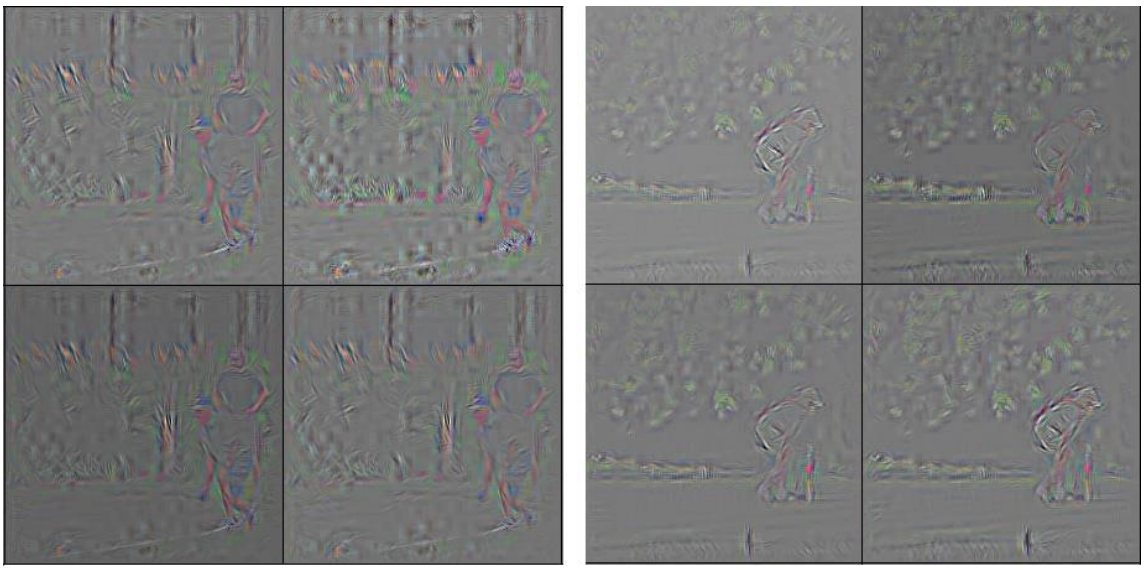
شکل ۴-۸ نمونه های استفاده شده برای نشان دادن مناسب نبودن الگوهای بصری در مرجع [۱۲۸]

در این قسمت، تفاوت الگوهای بصری استخراجی از کل تصویر در روش پیشنهادی رساله و مرجع [۱۲۸] نشان داده می‌شود. در روش پیشنهادی از ویژگی‌های استخراجی از اشیاء موجود در تصویر برای دسته‌بندی استفاده شد. در واقع ایتدا نواحی مرتبط به اشیاء استخراج و در گام بعدی برچسب آنها تشخیص داده شد. شکل ۴-۹ خروجی حاصل از نواحی اشیاء موجود در تصاویر شکل ۴-۸ را نشان می‌دهد. همانطور که نشان داده شد روش پیشنهادی عملکرد مناسبی در استخراج نواحی دارد. دو تصویر خروجی بعد از اعمال SVMفازی دو کلاسه در روش پیشنهادی می‌باشند.



شکل ۴-۹ اشیاء و نواحی استحراجی در دو تصویر مجموعه داده UIUC sport-8

شکل ۴-۱۰ نمونه‌هایی از خروجی لایه آخر پیچش معماری ResNet در دو تصویر شکل ۴-۹ را نشان می‌دهد. در این شکل کل تصاویر به عنوان ورودی در نظر گرفته شده است. برای نمایش این ویژگی‌ها از مفهوم واپیچش[۱]معرفی شده در مرجع [۱۴۵] استفاده شد. واپیچش روشی برای نمایش ویژگی‌های بصری در لایه‌های پیچش می‌باشد. همانطور که در شکل مشخص می‌باشد این لایه ساختار خوبی از کل تصویر ارائه داده است ولی نمی‌تواند تمایز بین اشیاء را به خوبی نشان دهد. از طرفی تاثیر پس‌زمینه در کل تصویر مشهود می‌باشد.

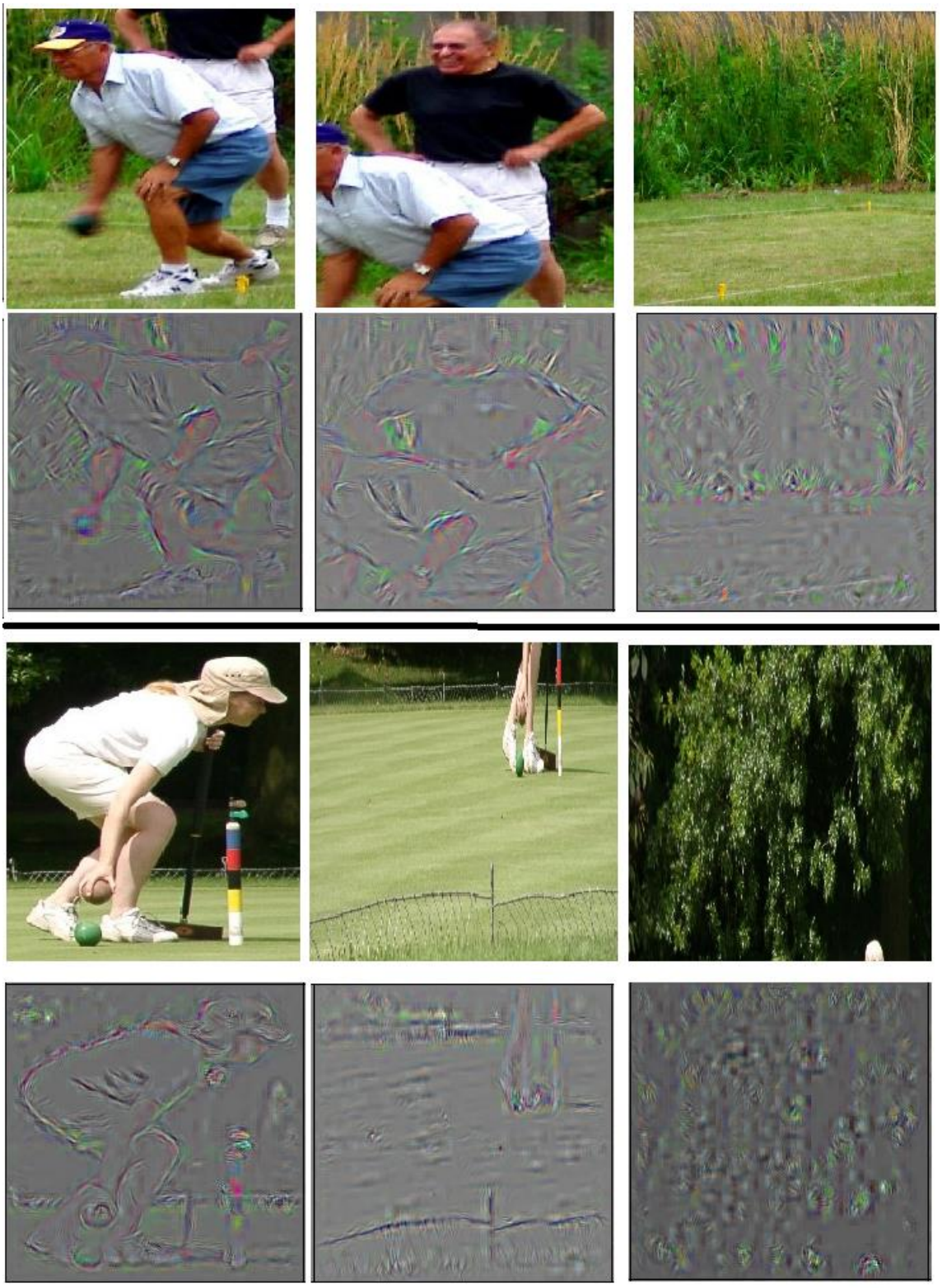


شکل ۴-۱۰ نمونه‌هایی از خروجی لایه آخر معماری ResNet توسط واپیچش با در نظر گرفتن کل تصویر

شکل ۴-۱۱ خروجی حاصل از واپیچش را برای اشیاء استخراجی توسط روش پیشنهادی را نشان می‌دهد. همانطور که در این شکل نشان داده شد، لایه آخر پیچش ResNet از لحاظ بصری ویژگی‌های سطح-بالاتری را برای هر شی به صورت جداگانه نسبت به کل تصویر استخراج می‌کند. در واقع این ویژگی بهتر می‌توانند تمایز بین صحنه ها را نشان دهند. تاثیر پس‌زمینه در دسته‌بندی کمتر شده است و اگر از نواحی مرتبط به اشیاء استفاده شود صحنه‌ها به راحتی می‌توانند از هم جدا شوند.

[۱]Deconvolution

با توجه به نتایج بدست آمده از این رساله می‌توان بیان نمود که استفاده از الگوهای بصری بهترین روش برای دسته‌بندی صحنه می‌باشند و حاشیه نویسی توسط کاربران وب می‌تواند با مشکلات زیادی روبرو شود.



شکل ۴-۱۱ نمونه هایی از خروجی لایه آخر معماری ResNet توسط واپیچش با درنظر گرفتن اشیاء به صورت جداگانه

۴ - ۷ جمع‌بندی

در این فصل سه روش برای دسته‌بندی تصاویر بر اساس پیچیدگی صحنه معرفی شده است. در روش اول، یک معماری جدید برای شبکه عصبی عمیق پیچش در دسته‌بندی تصاویر نامتعارف با دو کلاس معرفی شد که نتایج به دست آمده نشان‌دهنده بهبود عملکرد از لحاظ سرعت و دقت می‌باشد. دو روش بعدی برای دسته‌بندی تصاویر با تعداد کلاس‌های بیشتر معرفی گردید. در روش اول از روش‌های سنتی یعنی روش‌های قبل از یادگیری عمیق استفاده شد. در این بخش روشی بر پایه Boosting و نواحی کاندیدا معرفی شد. نتایج آزمایش‌های انجام شده نشان داد که این روش نسبت به روش‌های موجود سنتی یادگیری ماشین عملکرد مناسبی دارد ولی پیچیدگی زمانی این روش بسیار بالا بوده که نمی‌توان در مجموعه داده‌های حجیم از آن استفاده نمود. در نهایت یک روش جامع بر اساس استخراج نواحی و برچسب اشیاء و همچنین ویژگی‌های موضوعی در تصویر پیشنهاد شد. نتایج بیانگر بهبود عملکرد روش پیشنهادی نسبت به روش‌های موجود در حیطه دسته‌بندی تصویر است و می‌توان از این روش برای دسته‌بندی تصاویر با تعداد بالا استفاده نمود.

۵ جمع بندی و کارهای آتی

۵ - ۱ مقدمه

در این فصل جمع‌بندی کارهای انجام شده و تحلیل نتایج بدست آمده در رساله ارائه می‌شود. در ادامه کارهای آتی در بهبود سیستم‌های دسته‌بندی صحنه پیشنهاد می‌شود.

۵ - ۲ جمع‌بندی پژوهش انجام شده در رساله

با توجه به افزایش روز افزون تصاویر دیجیتال دسته‌بندی صحنه تصویر بر اساس محتوا به یک نیاز اساسی در سال‌های اخیر تبدیل شده است.. در همین راستا در این رساله چندین روش برای دسته‌بندی صحنه تصاویر پیشنهاد شده است. در این رساله، دو دیدگاه متفاوت برای مسئله دسته‌بندی صحنه تصویر در نظر گرفته شد. در دیدگاه نخست، تصاویر دو کلاسه مورد ارزیابی قرار گرفت. درواقع در این نوع از تصاویر بدست آوردن یک موضوع سراسری خاص مد نظر می‌باشد. از آنجایی که دسته‌بندی و تشخیص تصاویر نامتعارف در تصاویر تحت وب از اهمیت ویژه‌ای برخوردار می‌باشند به عنوان یک کاربرد مهم دسته‌بندی این نوع از تصاویر مورد ارزیابی قرار گرفت. در همین راستا روشی برای تشخیص و تمایز تصاویر نامتعارف بر پایه شبکه عصبی عمیق پیچش ارائه شد. در این روش یک معماری جدید از ترکیب دو معماری AlexNet و LeNet پیشنهاد شد. روش پیشنهادی بر روی مجموعه داده استاندارد با دو نوع توزیع متعادل و نامتعادل مورد ارزیابی قرار گرفت. نتایج بدست آمده در این بخش نشان دهنده بهبود عملکرد روش پیشنهادی نسبت به روش‌های ارائه شده در سال‌های اخیر می‌باشد.

در دیدگاه دوم، تصاویر با کلاس‌های بیشتر در نظر گرفته شده است. در واقع هر تصویر دارای یک موضوعی سراسری می‌باشد. در این دیدگاه دو روش ارائه شده است. در روش اول، روشی برپایه روش‌های سنتی یادگیری ماشین ارائه شد. در این روش ترکیبی از روش استخراج نواحی کاندیدا و Boosting پیشنهاد شده است. در گام نخست، از روش گروه‌بندی ترکیبی بر پایه چندین مقیاس که یک الگوریتم سریع برای محاسبه قطعه‌بندی سلسله مراتبی چندین مقیاسه می‌باشد برای استخراج نواحی کاندیدا

استفاده شده است. دلیل انتخاب این روش سرعت بالا و دقت مناسب بدست آمده از نتایج این رساله بر روی پایگاه داده‌های متفاوت بوده است. در گام بعدی از Boosting با ترکیب سه رگرسیون برای تشخیص برچسب نواحی کاندیدا استفاده شد. سپس توسط پیچش نواحی کاندیدا و SVD ویژگی استخراج شد. در نهایت توسط ویژگی‌های استخراجی و برچسب هر ناحیه کاندیدا دسته‌بندی صحنه توسط یک SVM انجام شده است. نتایج بدست آمده نشان دهنده عملکرد مناسب می‌باشد ولی مشکل این روش پیچیدگی زمانی آن می‌باشد. در واقع سرعت این روش در داده‌های حجیم با تعداد کلاس زیاد مناسب نمی‌باشد. با توجه به مطالعات و پیاده‌سازی‌های انجام شده روش‌های سنتی یادگیری ماشین نمی‌توانند به تنهایی در داده‌های حجیم عملکرد مناسبی را از خود نشان دهند. لازم به ذکر است که منظور از روش‌های سنتی، روش‌های ارائه شده قبل از یادگیری عمیق می‌باشد.

در روش دوم، یک سیستم جامع بر ساس ترکیب یادگیری عمیق و مدل‌های موضوعی ارائه شده است. این روش شامل چندین فاز مهم می‌باشد. در فاز اول، نواحی کاندیدا و برچسب نواحی استخراج شده است. برای استخراج نواحی کاندیدا، از شبکه عصبی عمیق بهبود داده شده R-FCN استفاده شد. در این قسمت برای محاسبه نواحی کاندیدا از شبکه ناحیه کاندیدای R-FCN با یک تابع زیان جدید استفاده شد. در گام بعدی با حذف لایه آخر این شبکه، یک معماری جدید بر اساس SVMفازی دو کلاسه و SVR برای تشخیص نواحی مناسبتر و برچسب‌زنی تصاویر پیشنهاد شده است. در روش پیشنهادی برای حذف نواحی کاندیدای نامناسب، یک SVM فازی دو کلاسه برای تمایز بین نواحی مرتبط به پس‌زمینه و شی استفاده شد. در واقع در این فاز بسیاری از نواحی کاندیدا حذف شد. این امر سبب بهبود سرعت با کاهش تعداد نواحی کاندیدا شد. در گام بعدی از SVR برای تشخیص برچسب نواحی استخراجی استفاده شد. برای بالا بردن سرعت پردازش، SVR ها به صورت موازی به ازای هر شی پیاده سازی شد. در فاز بعدی، از نواحی استخراجی SVMفازی دو کلاسه برای استخراج کلمات بصری استفاده شده است. در روش‌های که تاکنون ارائه شده از کل تصویر برای استخراج کلمات بصری استفاده می‌شد ولی در روش پیشنهادی از نواحی کاندیدای استخراجی توسط SVM فازی دو کلاسه

استفاده شده است. این امر سبب بالا بردن دقت استخراج کلمات بصری می‌شود. در ادامه برای استخراج مجموعه کلمات بصری از یک k-means توسعه یافته برای بالا بردن سرعت عملکرد این روش استفاده شد. سپس از کلمات بصری به همراه برچسب‌های استخراجی مرتبط به ناحیه آنها برای محاسبه ویژگی‌های موضوعی استفاده شد. برای محاسبه ویژگی‌های موضوعی روش‌های بسیاری بر پایه LDA مورد ارزیابی قرار گرفت ولی پیچیدگی زمانی این روش‌ها بسیار بالا بوده است. در همین راستا از یک شبکه عمیق با عنوان تخمین‌گر توزیع‌شده خودرگرسیو عصبی اسناد استفاده شد که زمان مرحله آزمون بسیار خوبی داشته است. این شبکه عمیق، برای دسته‌بندی تصویر بر اساس ویژگی‌های موضوع معرفی شد که در این رساله فقط از لایه مرتبط به ویژگی‌های موضوعی استفاده شده است. لازم به ذکر است در روش تخمین‌گر توزیع شده خودرگرسیو عصبی اسناد برچسب نواحی به عنوان ورودی سیستم بوده که در روش پیشنهادی به صورت خودکار توسط مرحله قبل استخراج شد. در نهایت در فاز آخر، از یک دسته‌بند بیزین بر پایه ویژگی‌های استخراج شده توسط یادگیری عمیق تمام پیچش، ویژگی‌های موضوعی و برچسب نواحی بصری برای دسته‌بندی صحنه تصویر استفاده شد.

روش پیشنهادی از لحاظ زمان و دقت بر روی مجموعه داده‌های Scene-15، UIUC Sports و MIT-67 به ترتیب با ۱۵، ۳۶ و ۶۷ دسته متفاوت مورد ارزیابی قرار گرفت. نتایج نشان دهنده بهبود عملکرد روش پیشنهادی از لحاظ دقت و زمان مرحله آزمون در این مجموعه داده‌ها می‌باشد.

نوآوری‌های ارائه شده در این رساله:

۱. ارائه یک شبکه عصبی عمیق پیچش برپایه معماری AlexNet و LeNet برای دسته‌بندی تصاویر دو کلاسه
۲. استفاده از روش‌های سنتی یادگیری ماشین بر پایه نواحی کاندیدا و Boosting برای برچسب-زنی و دسته‌بندی تصویر
۳. بهبود R-FCN برای تشخیص نواحی کاندیدا و برچسب شی

a. استفاده از تابع زیان جدید
b. استفاده از SVMفازی دو کلاسه برای کاهش نواحی کاندیدا و تفکیک بین این نواحی از پس‌زمینه
c. استفاده از SVR برای برچسب زنی تصاویر

۴. استخراج کلمات کلیدی بر اساس نواحی استخراج توسط SVMفازی دو کلاسه و استفاده از k-means برای اسخراج ویژگی سرآمد
۵. استفاده خودکار از کلمات کلیدی و برچسب نواحی برای استخراج ویژگی‌های موضوعی
۶. استفاده از تخمین‌گر توزیع‌شده خودرگرسیو عصبی اسناد برپایه کلمات بصری استخراجی و موقعیت کلمات بصری
۷. ترکیب ویژگی‌های استخراجی از لایه آخر شبکه عصبی عمیق تمام پیچش، برچسب کلمات کلیدی بر پایه ناحیه قرار گرفتن آنها و ویژگی‌های موضوعی برای دسته‌بندی صحنه
۸. دسته‌بندی صحنه توسط بیزین بر اساس ویژگی‌های استخراجی

۵ - ۳ کارهای آینده

در این پژوهش راهکاری نوین برای دسته‌بندی صحنه تصویر ارائه شده است. در راهکار ارائه شده تمرکز بسیاری بر استخراج نواحی مرتبط به اشیاء، برچسب اشیاء، به دست آوردن موضوعات پنهان در تصویر و استخراج ویژگی‌های سطح بالا بر اساس شبکه‌های عمیق تمام پیچش شده است. در این بخش، پیشنهادهایی برای کارهای آینده آورده می‌شود.

۱. یک فاز بسیار مهم در سیستم‌های تشخیص اشیاء، به دست آوردن نواحی کاندیدا می‌باشد که دقت این بخش در مرحله بعد یعنی تحلیل سیستم‌های شناسایی و آشکارسازی شی، بسیار تاثیرگذار می‌باشد. در این رساله از شبکه تمام پیچش بر پایه معماری ResNet، ZF و VGGNet برای محاسبه نواحی کاندیدا استفاده شده است. با توجه به نتایج به دست آمده با ا

معماری‌های حاضر نمی‌توان بهبود بیشتری را در استخراج نواحی کاندیدا بدست آورد. از این‌رو ارائه یک معماری جدید با ساختار نوین می‌تواند در دقت روش پیشنهادی بسیار تاثیرگذار باشد.

۲. در این رساله برای محاسبه مجموعه کلمات از روش کلمات بصری بر پایه ویژگی SIFT و خوشه‌بندی k-means توسعه یافته استفاده شد. این روش عملکرد مناسبی در استخراج مجموعه کلمات از خود نشان داد. طبق مشاهده‌های عملی در هنگام دریافت تصاویر در مغز تنها قسمت کمی از نورون‌ها درگیر هستند. لذا می‌توان دریافت، نمایش سطح بالاتر و احتمالاً تنک در این فرآیند دخیل هست. علاوه براین، آزمایش‌های انجام شده در این رساله نشان‌دهنده تنک بودن ماتریس استخراجی از کلمات بصری می‌باشد. در واقع می‌توان برای این نوع از سیستم‌ها از روش کدگذاری تنک با تغییر مکانیزم کدگذاری اولیه استفاده نمود. کدگذاری انجام شده در این رساله، مستقل از تنک بودن تکه‌های موجود در تصویر بوده است ولی در نظر گرفتن این واقعیت و اعمال آن در کدگذاری، خطای بازسازی را کاهش خواهد داد. علاوه بر این، در بسیاری از کارهای انجام شده از کدگذاری مستقل از محتوای تصویر استفاده شده است. به عبارت دیگر تمام تکه‌های تصویر به یک شکل کدگذاری و در طبقه‌بندی لحاظ شده‌اند. البته در هنگام تجمیع با تابع بیشینه تاثیر بعضی از تکه‌ها کمتر می‌شود. با این حال بهتر است مرحله کدگذاری با توجه به محتوای تصاویر انجام شود تا اطلاعات اضافی استخراج نشود و تکه‌های کدگذاری شده مناسب طبقه‌بندی باشند.

۳. استخراج موضوعات نهفته در تصویر می‌تواند در بالا بردن دقت عملکرد سیستم‌های دسته‌بندی تصور بسیار تاثیرگذار باشند. با توجه به روش‌های معرفی شده در سال‌های اخیر و آزمایش‌های انجام شده، این روش‌ها دقت خوبی در مقایسه با روش‌های پیشین از خود نشان داده‌اند. ولی پیچیدگی زمانی این نوع از روش‌ها بسیار بالا است. در همین راستا معرفی و به کارگیری روشی نوین بر پایه مدل‌های موضوعی که زمان اجرای بهتری دارند، می‌تواند در این کاربرد بسیار با اهمیت باشند.

۴. یکی از کاربردهای مهم در بینایی ماشین استفاده از سیستم‌هایی بر پایه محتوا برای توصیف خودکار تصویر می باشند. همانطور که انسان بعد از تحلیل محتوای تصویر به توصیف خودکار آن می‌پردازد، می‌تواند از سیستم ارائه شده در این رساله برای گام بعدی مانند توصیف خودکار تصویر استفاده نمود. در واقع می‌توان از این روش در سیستم‌های بازیابی مبتنی بر معنا[۱] (SBIR) استفاده کرد. در SBIR، کاربر، معنای مورد نظر خود را به صورت یک عبارت متنی وارد کرده و به دنبال یافتن تصاویری با محتویات مرتبط به آن عبارت است. در واقع در این سامانه، رابطه بین معنا و محتوای بصری تصاویر بررسی می‌شود. منطقی‌ترین روش برای جستجو سطح بالای معنایی در سامانه‌های SBIR، انتساب برچسب متنی مبتنی بر معنا به تصاویر موجود در پایگاه داده و مقایسه بین این برچسب‌ها و عبارت مورد جستجو برای بازیابی تصاویر تصاویر مربوطه می‌باشد. در ادامه این رساله، می‌توان از خروجی روش ارائه شده برای برچشب‌زنی در سامانه SBIR استفاده نمود. یعنی برچسب‌های محتوایی از تصویر توسط روش پیشنهادی رساله استخراج و به عنوان ورودی به سامانه SBIR اعمال شوند.

۵. در این رساله از مجموعه داده‌هایی با چند شی استفاده شده است. عملکرد سیستم پیشنهادی بر نواحی استخراجی به این اشیاء بسیار وابسته است. در صورتی که این اشیاء به خوبی استخراج نشوند عملکرد این نوع از سیستم‌ها بسیار کاهش می‌یابد. در واقع از نواحی مرتبط به این اشیاء کلمات بصری استخراج می‌شود. در سیستم‌هایی که تاکنون برای دسته‌بندی محتوایی تصویر ارائه شده است از کل تصویر برای استخراج کلمات بصری استفاده شد. در واقع در این روش‌ها کلمات بصری برپایه نواحی محلی که از کل تصویر استخراج شده‌اند، محاسبه می‌شوند. می‌توان سیستمی ترکیبی از سیستم پیشنهادی رساله و سیستم نوین برپایه استخراج نواحی محلی که

[۱] Semantic Based Image Retrieval (SBIR)

وابسته به برچسب اشیاء نمی‌باشد ارائه نمود و از سیستمی ترکیبی برای دسته‌بندی صحنه تصویر ارائه کرد.

منابع

[1] Y. Zheng, Y.-J. Zhang, and H. Larochelle, "A deep and autoregressive approach for topic modeling of multimodal data," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, pp. 1056-1069, 2016.

[2] M. Zang, D. Wen, K. Wang, T. Liu, and W. Song, "A novel topic feature for image scene classification," *Neurocomputing*, vol. 148, pp. 467-476, 2015.

[3] W. Chong, D. Blei, and F.-F. Li, "Simultaneous image classification and annotation," in *Computer Vision and Pattern Recognition. IEEE Conference on*, 2009, pp. 1903-1910.

[4] Z. Niu, G. Hua, X. Gao, and Q. Tian, "Context aware topic model for scene recognition," in *Computer Vision and Pattern Recognition, IEEE Conference on*, 2012, pp. 2743-2750.

[5] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in *Workshop on statistical learning in computer vision, ,* 2004, pp. 1-2.

[6] E. Nowak, F. Jurie, and B. Triggs, "Sampling strategies for bag-of-features image classification," in *European conference on computer vision*, 2006, pp. 490-503.

[7] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *Computer Vision and Pattern Recognition. IEEE Conference on*, 2009, pp. 1794-1801.

[8] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition,*, 2006, pp. 2169-2178.

[9] K. Grauman and T. Darrell, "The pyramid match kernel: Discriminative classification with sets of image features," in *Tenth IEEE International Conference on Computer Vision, Volume 1*, 2005, pp. 1458-1465.

[10] K. Yu, T. Zhang, and Y. Gong, "Nonlinear learning using local coordinate coding," in *Advances in neural information processing systems*, 2009, pp. 2223-2231.

[11] K. Sjöstrand, "Matlab implementation of LASSO, LARS, the elastic net and SPCA," *Informatics and Mathematical Modelling, Technical University of Denmark*, 2005.

[12] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in *Computer Vision and Pattern Recognition, IEEE Conference on*, 2010, pp. 3360-3367.

[13] J. C. Van Gemert, C. J. Veenman, A. W. Smeulders, and J.-M. Geusebroek, "Visual word ambiguity," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, pp. 1271-1283, 2009.

[14] L. Liu, L. Wang, and X. Liu, "In defense of soft-assignment coding," in *International Conference on Computer Vision*, 2011, pp. 2486-2493.

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* vol. 37, pp. 1904-1916, 2015.

[16] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1440-1448.

[17] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems*, 2015, pp. 91-99.

[18] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580-587.

[19] J. R. Uijlings, K. E. van de Sande, T. Gevers, and A. W. Smeulders, "Selective search for object recognition," *International journal of computer vision*, vol. 104, pp. 154-171, 2013.

[20] C. L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges," in *European Conference on Computer Vision*, 2014, pp. 391-405.

[21] K. E. Van de Sande, J. R. Uijlings, T. Gevers, and A. W. Smeulders, "Segmentation as selective search for object recognition," in *Computer Vision, IEEE International Conference on*, 2011, pp. 1879-1886.

[22] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik, "Contour detection and hierarchical image segmentation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* vol. 33, pp. 898-916, 2011.

[23] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient graph-based image segmentation," *International Journal of Computer Vision,* vol. 59, pp. 167-181, 2004.
[24] S. Manen, M. Guillaumin, and L. Gool, "Prime object proposals with randomized prim's algorithm," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 2536-2543.
[25] P. Rantalankila, J. Kannala, and E. Rahtu, "Generating object segmentation proposals using global and local search," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 2417-2424.
[26] K.-Y. Chang, T.-L. Liu, H.-T. Chen, and S.-H. Lai, "Fusing generic objectness and visual saliency for salient object detection," in *Computer Vision, IEEE International Conference on*, 2011, pp. 914-921.
[27] J. Carreira and C. Sminchisescu, "Constrained parametric min-cuts for automatic object segmentation," in *Computer Vision and Pattern Recognition, IEEE Conference on*, 2010, pp. 3241-3248.
[28] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International journal of computer vision,* vol. 88, pp. 303-338, 2010.
[29] I. Endres and D. Hoiem, "Category-independent object proposals with diverse ranking," *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* vol. 36, pp. 222-234, 2014.
[30] D. Hoiem, A. A. Efros, and M. Hebert, "Recovering occlusion boundaries from an image," *International Journal of Computer Vision,* vol. 91, pp. 328-346, 2011.
[31] S. Gupta, R. Girshick, P. Arbeláez, and J. Malik, "Learning rich features from RGB-D images for object detection and segmentation," in *European Conference on Computer Vision*, 2014, pp. 345-360.
[32] A. Humayun, F. Li, and J. Rehg, "RIGOR: reusing inference in graph cuts for generating object regions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 336-343.
[33] P. Krähenbühl and V. Koltun, "Geodesic object proposals," in *Computer Vision*, ed, 2014, pp. 725-739.
[34] P. Dollár and C. L. Zitnick, "Fast edge detection using structured forests," *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* vol. 37, pp. 1558-1570, 2015.
[35] P. Arbeláez, J. Pont-Tuset, J. T. Barron, F. Marques, and J. Malik, "Multiscale combinatorial grouping," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 328-335.
[36] B. Alexe, T. Deselaers, and V. Ferrari, "What is an object?," in *Computer Vision and Pattern Recognition, IEEE Conference on*, 2010, pp. 73-80.
[37] E. Rahtu, J. Kannala, and M. Blaschko, "Learning a category independent object detection cascade," in *Computer Vision, IEEE International Conference on*, 2011, pp. 1052-1059.
[38] M.-M. Cheng, Z. Zhang, W.-Y. Lin, and P. Torr, "BING: Binarized normed gradients for objectness estimation at 300fps," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 3286-3293.
[39] J. Kim and K. Grauman, "Shape sharing for object segmentation," in *Computer Vision*, ed: Springer, 2012, pp. 444-458.
[40] D. Erhan, C. Szegedy, A. Toshev, and D. Anguelov, "Scalable object detection using deep neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 2147-2154.
[41] J. Yan, Y. Yu, X. Zhu, Z. Lei, and S. Z. Li, "Object detection by labeling superpixels," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 5107-5116.
[42] ر. راد و م. جم زاد، "مروری بر سیستم های برچسب زنی تصاویر،" مجله ماشین بینایی و پردازش تصویر، آماده برای انتشار، سال ۱۳۹۸.
[43] ر. راد، "برچسب زنی خودکار تصاویر بر مبنای تجزیه ماتریس نامنفی به صورت چند منظری" رساله دکتری، دانشکده کامپیوتر، دانشگاه صنعتی شریف، سال ۱۳۹۶.
[44] س. ش. بهائی، "بازشناسی اشیاء در تصاویر دیجیتال با استفاده از روش های مبتنی بر متن" پایان نامه کارشناسی ارشد، دانشکده فنی و مهندسی، دانشگاه خوارزمی تهران، سال ۱۳۹۳.
[45] K. Lenc and A. Vedaldi, "R-cnn minus r," *arXiv preprint arXiv:1506.06981*, 2015.

[46] C. Sun, M. Paluri, R. Collobert, R. Nevatia, and L. Bourdev, "Pronet: Learning to propose object-specific boxes for cascaded neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3485-3493.
[47] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, *et al.*, "Ssd: Single shot multibox detector," in *European conference on computer vision*, 2016, pp. 21-37.
[48] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 779-788.
[49] M. Najibi, M. Rastegari, and L. S. Davis, "G-cnn: an iterative grid based object detector," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2369-2377.
[50] C.-Y. Fu, W. Liu, A. Ranga, A. Tyagi, and A. C. Berg, "DSSD: Deconvolutional Single Shot Detector," *arXiv preprint arXiv:1701.06659*, 2017.
[51] J. Redmon and A. Farhadi, "YOLO9000: better, faster, stronger," *arXiv preprint arXiv:1612.08242*, 2016.
[52] J. Dai, Y. Li, K. He, and J. Sun, "R-fcn: Object detection via region-based fully convolutional networks," in *Advances in neural information processing systems*, 2016, pp. 379-387.
[53] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770-778.
[54] R. Richmond, "Facebook's new way to combat child pornography," *New York Times*, 2011.
[55] N. M. Malamuth, "Criminal and noncriminal sexual aggressors," *Annals of the New York Academy of Sciences*, vol. 989, pp. 33-58, 2003.
[56] E. M. Alexy, A. W. Burgess, and R. A. Prentky, "Pornography use as a risk marker for an aggressive pattern of behavior among sexually reactive children and adolescents," *Journal of the American Psychiatric Nurses Association*, vol. 14, pp. 442-453, 2009.
[57] "http://www.dailyinfographic.com/the-stats-on-internet-pornography-nfographic,accessed, ," *Available: 2017/1/19*.
[58] T. Largillier, G. Peyronnet, and S. Peyronnet, "Efficient filtering of adult content using textual information," *arXiv preprint arXiv:1512.00198*, pp. 14-17, 2016.
[59] D. A. Forsyth and M. M. Fleck, "Identifying nude pictures," in *Applications of Computer Vision. Proceedings 3rd IEEE Workshop on*, 1996, pp. 103-108.
[60] W. Hu, O. Wu, Z. Chen, Z. Fu, and S. Maybank, "Recognition of pornographic web pages by classifying texts and images," *IEEE transactions on pattern analysis and machine intelligence*, vol. 29, pp. 1019-1034, 2007.
[61] J. Zhang, L. Sui, L. Zhuo, Z. Li, and Y. Yang, "An approach of bag-of-words based on visual attention model for pornographic images recognition in compressed domain," *Neurocomputing*, vol. 110, pp. 145-152, 2013.
[62] L. Sui, J. Zhang, L. Zhuo, and Y. Yang, "Research on pornographic images recognition method based on visual words in a compressed domain," *IET image processing*, vol. 6, pp. 87-93, 2012.
[63] K. Dong, L. Guo, and Q. Fu, "An adult image detection algorithm based on Bag-of-Visual-Words and text information," in *10th International Conference on Natural Computation* 2014, pp. 556-560.
[64] C. X. Ries and R. Lienhart, "A survey on visual adult image recognition," *Multimedia tools and applications*, vol. 69, pp. 661-688, 2014.
[65] D. Li, N. Li, J. Wang, and T. Zhu, "Pornographic images recognition based on spatial pyramid partition and multi-instance ensemble learning," *Knowledge-Based Systems*, vol. 84, pp. 214-223, 2015.
[66] R. Lienhart and R. Hauke, "Filtering adult image content with topic models," in *IEEE International Conference on Multimedia and Expo*, 2009, pp. 1472-1475.
[67] Y.-L. Boureau, F. Bach, Y. LeCun, and J. Ponce, "Learning mid-level features for recognition," in *Computer Vision and Pattern Recognition, IEEE Conference on*, 2010, pp. 2559-2566.
[68] L.-J. Li, H. Su, L. Fei-Fei, and E. P. Xing, "Object bank: A high-level image representation for scene classification & semantic feature sparsification," in *Advances in neural information processing systems*, 2010, pp. 1378-1386.
[69] T. Hofmann, "Probabilistic latent semantic indexing," in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, 1999, pp. 50-57.

[70] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research,* vol. 3, pp. 993-1022, 2003.
[71] L. Fei-Fei and P. Perona, "A bayesian hierarchical model for learning natural scene categories," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2005, pp. 524-531.
[72] E. B. Sudderth, A. Torralba, W. T. Freeman, and A. S. Willsky, "Learning hierarchical models of scenes, objects, and parts," in *Tenth IEEE International Conference on Computer Vision, Volume 1*, 2005, pp. 1331-1338.
[73] Z. Niu, G. Hua, X. Gao, and Q. Tian, "Spatial-DiscLDA for visual recognition," in *Computer Vision and Pattern Recognition, IEEE Conference on*, 2011, pp. 1769-1776.
[74] S. Lacoste-Julien, F. Sha, and M. I. Jordan, "DiscLDA: Discriminative learning for dimensionality reduction and classification," in *Advances in neural information processing systems*, 2009, pp. 897-904.
[75] J. Zhu, L.-J. Li, L. Fei-Fei, and E. P. Xing, "Large margin learning of upstream scene understanding models," in *Advances in Neural Information Processing Systems*, 2010, pp. 2586-2594.
[76] Y. Su and F. Jurie, "Visual word disambiguation by semantic contexts," in *International Conference on Computer Vision*, 2011, pp. 311-318.
[77] L.-J. Li and L. Fei-Fei, "What, where and who? classifying events by scene and object recognition," in *IEEE 11th International Conference on Computer Vision*, 2007, pp. 1-8.
[78] R. Fergus, P. Perona, and A. Zisserman, "Object class recognition by unsupervised scale-invariant learning," in *Computer Vision and Pattern Recognition, Proceedings. IEEE Computer Society Conference on*, 2003, pp. 264-271.
[79] T. Harada, Y. Ushiku, Y. Yamashita, and Y. Kuniyoshi, "Discriminative spatial pyramid," in *Computer Vision and Pattern Recognition, IEEE Conference on*, 2011, pp. 1617-1624.
[80] Y. Cao, C. Wang, Z. Li, L. Zhang, and L. Zhang, "Spatial-bag-of-features," in *Computer Vision and Pattern Recognition, IEEE Conference on*, 2010, pp. 3352-3359.
[81] X. Wang and E. Grimson, "Spatial latent dirichlet allocation," in *Advances in neural information processing systems*, 2008, pp. 1577-1584.
[82] D. Liu, G. Hua, P. Viola, and T. Chen, "Integrated feature selection and higher-order spatial feature extraction for object categorization," in *Computer Vision and Pattern Recognition. IEEE Conference on*, 2008, pp. 1-8.
[83] L. Cao and L. Fei-Fei, "Spatially coherent latent topic model for concurrent segmentation and classification of objects and scenes," in *IEEE 11th International Conference on Computer Vision*, 2007, pp. 1-8.
[84] H. Larochelle and S. Lauly, "A neural autoregressive topic model," in *Advances in Neural Information Processing Systems*, 2012, pp. 2708-2716.
[85] G. Heinrich, "Parameter estimation for text analysis," *University of Leipzig, Tech. Rep*, 2008.
[86] C. M. Bishop and N. M. Nasrabadi, "Pattern recognition and machine learning.," vol. 1, ed: Springer, New York, 2006.
[87] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097-1105.
[88] Y. Bengio, "Learning deep architectures for AI," *Foundations and trends in Machine Learning*, vol. 2, pp. 1-127, 2009.
[89] Q. V. Le, "Building high-level features using large scale unsupervised learning," in *Acoustics, Speech and Signal Processing, IEEE International Conference on*, 2013, pp. 8595-8598.
[90] R. T. d. Combes, M. Pezeshki, S. Shabanian, A. Courville, and Y. Bengio, "On the Learning Dynamics of Deep Neural Networks," *arXiv preprint arXiv:1809.06848*, 2018.
[91] K. Janocha and W. M. Czarnecki, "On Loss Functions for Deep Neural Networks in Classification," *arXiv preprint arXiv:1702.05659*, 2017.
[92] B. Fasel, "Robust face analysis using convolutional neural networks," in *Pattern Recognition, 16th International Conference on*, 2002, pp. 40-43.
[93] Y. LeCun and Y. Bengio, "Convolutional networks for images, speech, and time series," *The handbook of brain theory and neural networks*, vol. 3361, 1995.
[94] R. E. Schapire, "The boosting approach to machine learning: An overview," in *Nonlinear estimation and classification*, ed: Springer, 2003, pp. 149-171.
[95] J. Friedman, T. Hastie, and R. Tibshirani, "Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors)," *The annals of statistics*, vol. 28, pp. 337-407, 2000.

[96] A. Torralba, K. P. Murphy, and W. T. Freeman, "Sharing visual features for multiclass and multiview object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, pp. 854-869, 2007.
[97] R. Lienhart, A. Kuranov, and V. Pisarevsky, "Empirical analysis of detection cascades of boosted classifiers for rapid object detection," *Pattern Recognition*, pp. 297-304, 2003.
[98] W. M. Czarnecki, R. Jozefowicz, and J. Tabor, "Maximum entropy linear manifold for learning discriminative low-dimensional representation," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 2015, pp. 52-67.
[99] S.-J. Yen, Y.-C. Wu, J.-C. Yang, Y.-S. Lee, C.-J. Lee, and J.-J. Liu, "A support vector machine-based context-ranking model for question answering," *Information Sciences*, vol. 224, pp. 77-87, 2013.
[100] S.-G. Chen and X.-J. Wu, "A new fuzzy twin support vector machine for pattern classification," *International Journal of Machine Learning and Cybernetics*, pp. 1-12, 2017.
[101] J. M. Keller and D. J. Hunt, "Incorporating fuzzy membership functions into the perceptron algorithm," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 693-699, 1985.
[102] X. Peng, L. Wang, X. Wang, and Y. Qiao, "Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice," *Computer Vision and Image Understanding*, vol. 150, pp. 109-125, 2016.
[103] H. Wang, M. M. Ullah, A. Klaser, I. Laptev, and C. Schmid, "Evaluation of local spatio-temporal features for action recognition," in *British Machine Vision Conference*, 2009, pp. 124.1-124.11.
[104] S. Na, L. Xumin, and G. Yong, "Research on k-means clustering algorithm: An improved k-means clustering algorithm," in *Intelligent Information Technology and Security Informatics, Third International Symposium on*, 2010, pp. 63-67.
[105] S. M. Kia, H. Rahmani, R. Mortezaei, M. E. Moghaddam, and A. Namazi, "A Novel Scheme for Intelligent Recognition of Pornographic Images," *arXiv preprint arXiv:1402.5792*, 2014.
[106] A. Ahmadi, M. Fotouhi, and M. Khaleghi, "Intelligent classification of web pages using contextual and visual features," *Applied Soft Computing*, vol. 11, pp. 1638-1647, 2011.
[107] Q.-F. Zheng, W. Zeng, W.-Q. Wang, and W. Gao, "Shape-based adult image detection," *International Journal of Image and Graphics*, vol. 6, pp. 115-124, 2006.
[108] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba, "Sun database: Large-scale scene recognition from abbey to zoo," in *Computer vision and pattern recognition, IEEE conference on*, 2010, pp. 3485-3492.
[109] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International journal of computer vision*, vol. 42, pp. 145-175, 2001.
[110] Z. Li, Z. Shi, X. Liu, and Z. Shi, "Modeling continuous visual features for semantic image annotation and retrieval," *Pattern Recognition Letters*, vol. 32, pp. 516-523, 2011.
[111] X. Wang, M. Yang, S. Zhu, and Y. Lin, "Regionlets for generic object detection," in *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 17-24.
[112] R. B. Girshick, P. F. Felzenszwalb, and D. McAllester, "Discriminatively trained deformable part models, release 5," 2012.
[113] S. N. Parizi, J. G. Oberlin, and P. F. Felzenszwalb, "Reconfigurable models for scene recognition," in *Computer Vision and Pattern Recognition, IEEE Conference on*, 2012, pp. 2775-2782.
[114] L.-J. Li, H. Su, Y. Lim, and F.-F. Li, "Objects as Attributes for Scene Classification," in *ECCV Workshops*, 2010, pp. 57-69.
[115] J. Wu and J. M. Rehg, "CENTRIST: A visual descriptor for scene categorization," *IEEE transactions on pattern analysis and machine intelligence*, vol. 33, pp. 1489-1501, 2011.
[116] Y. Jiang, J. Yuan, and G. Yu, "Randomized spatial partition for scene recognition," in *Computer Vision*, ed: Springer, 2012, pp. 730-743.
[117] R. Kwitt, N. Vasconcelos, and N. Rasiwasia, "Scene recognition on the semantic manifold," in *European Conference on Computer Vision*, 2012, pp. 359-372.
[118] L. Torresani, M. Szummer, and A. Fitzgibbon, "Efficient object category recognition using classemes," in *European conference on computer vision*, 2010, pp. 776-789.
[119] J. Wu and J. M. Rehg, "Beyond the euclidean distance: Creating effective visual codebooks using the histogram intersection kernel," in *Computer Vision, IEEE 12th International Conference on*, 2009, pp. 630-637.
[120] S. Gao, I. W.-H. Tsang, L.-T. Chia, and P. Zhao, "Local features are not lonely–Laplacian sparse coding for image classification," in *Computer Vision and Pattern Recognition, IEEE Conference on*, 2010, pp. 3555-3561.

[121] F. Sadeghi and M. F. Tappen, "Latent pyramidal regions for recognizing scenes," in *European Conference on Computer Vision*, 2012, pp. 228-241.
[122] Y. Zheng, Y.-G. Jiang, and X. Xue, "Learning hybrid part filters for scene recognition," in *European Conference on Computer Vision*, 2012, pp. 172-185.
[123] A. Shabou and H. LeBorgne, "Locality-constrained and spatially regularized coding for scene categorization," in *Computer Vision and Pattern Recognition, IEEE Conference on*, 2012, pp. 3618-3625.
[124] Q. Li, J. Wu, and Z. Tu, "Harvesting mid-level visual concepts from large-scale internet images," in *Computer Vision and Pattern Recognition, IEEE Conference on*, 2013, pp. 851-858.
[125] A. Vedaldi and B. Fulkerson, "VLFeat: An open and portable library of computer vision algorithms," in *Proceedings of the 18th ACM international conference on Multimedia*, 2010, pp. 1469-1472.
[126] D. Lin, C. Lu, R. Liao, and J. Jia, "Learning important spatial pooling regions for scene classification," in *Computer Vision and Pattern Recognition, IEEE Conference on*, 2014, pp. 3726-3733.
[127] S. H. Khan, M. Hayat, M. Bennamoun, R. Togneri, and F. A. Sohel, "A discriminative representation of convolutional features for indoor scene recognition," *IEEE Transactions on Image Processing*, vol. 25, pp. 3372-3383, 2016.
[128] D. Wang and K. Mao, "Task-generic semantic convolutional neural network for web text-aided image classification," *Neurocomputing*, vol. 329, pp. 103-115, 2019.
[129] D. Wang and K. Mao, "Learning semantic text features for web text aided image classification," *major revised for IEEE Transactions on Multimedia*.
[130] A. Bosch, A. Zisserman, and X. Muñoz, "Scene classification using a hybrid generative/discriminative approach," *IEEE transactions on pattern analysis and machine intelligence*, vol. 30, pp. 712-727, 2008.
[131] R. Margolin, L. Zelnik-Manor, and A. Tal, "Otc: A novel local descriptor for scene classification," in *European Conference on Computer Vision*, 2014, pp. 377-391.
[132] J. Yuan, M. Yang, and Y. Wu, "Mining discriminative co-occurrence patterns for visual recognition," in *Computer Vision and Pattern Recognition, IEEE Conference on*, 2011, pp. 2777-2784.
[133] A. Quattoni and A. Torralba, "Recognizing indoor scenes," in *Computer Vision and Pattern Recognition. IEEE Conference on*, 2009, pp. 413-420.
[134] M. Pandey and S. Lazebnik, "Scene recognition and weakly supervised object localization with deformable part-based models," in *Computer Vision, IEEE International Conference on*, 2011, pp. 1307-1314.
[135] M. Juneja, A. Vedaldi, C. Jawahar, and A. Zisserman, "Blocks that shout: Distinctive parts for scene classification," in *Computer Vision and Pattern Recognition, IEEE Conference on*, 2013, pp. 923-930.
[136] S. Singh, A. Gupta, and A. A. Efros, "Unsupervised discovery of mid-level discriminative patches," in *Computer Vision*, ed: Springer, 2012, pp. 73-86.
[137] J. Sun and J. Ponce, "Learning discriminative part detectors for image classification and cosegmentation," in *Computer Vision, IEEE International Conference on*, 2013, pp. 3400-3407.
[138] C. Doersch, A. Gupta, and A. A. Efros, "Mid-level visual element discovery as discriminative mode seeking," in *Advances in neural information processing systems*, 2013, pp. 494-502.
[139] Y. Gong, L. Wang, R. Guo, and S. Lazebnik, "Multi-scale orderless pooling of deep convolutional activation features," in *European conference on computer vision*, 2014, pp. 392-407.
[140] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN features off-the-shelf: an astounding baseline for recognition," in *Computer Vision and Pattern Recognition Workshops, IEEE Conference on*, 2014, pp. 512-519.
[141] M. Cimpoi, S. Maji, and A. Vedaldi, "Deep filter banks for texture recognition and segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3828-3836.
[142] A. Diba, A. M. Pazandeh, and L. Van Gool, "Deep visual words: Improved fisher vector for image classification," in *Proceedings of the fifteenth IAPR International Conference on Machine Vision Applications MVA2017*, 2017, pp. 186-189.
[143] D. Yoo, S. Park, J.-Y. Lee, and I. So Kweon, "Multi-scale pyramid pooling for deep convolutional representation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2015, pp. 71-80.

[144] S. Gupta, D. K. Pradhan, D. A. Dinesh, and V. Thenkanidiyoor, "Deep Spatial Pyramid Match Kernel for Scene Classification," *In Proceedings of the 7th International Conference on Pattern Recognition Applications and Methods, 2018,* vol. 1, pp. 141-148, 2018.
[145] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European conference on computer vision*, 2014, pp. 818-833.

Abstract

Considering the fact that digital images are all over the internet and shared in the media, image classification has been recognized as one of the basic needs of digital world. Automatic content description is one of the major issues of image classification which has resulted in establishing a higher relationship between machine vision and image processing. Automatic content description has many applications in the web-based systems and search engines such as filtering unconventional images, identification of subjective images and detection of human behaviors.

Automatic content description problem can be studied from two points of view. In the first viewpoint, dataset includes only small images with specific applications such as recognition of unconventional images. Recognition of unconventional images is a description problem with two classes and minor complexities. In this thesis, a new structure for deep neural networks was proposed for detection of unconventional images. This new structure works principally based on the extraction of high-level features from human body in unconventional images. The results obtained from applying this new approach on two datasets indicated that the performance can be improved by 4% compared to other methods presented in the recent years.

In the second viewpoint, the dataset includes images with more classes of different scenes. In this thesis, an approach was proposed for an automatic content description based on the extraction of high-level semantic regions which was applied to obtain the label of objects. This method consists of several steps. In this first step, regions and label of objects are identified. To do this, a new structure based on the deep neural network was proposed which employs the R-FCN method. A loss function which has not been used before was also part of this method. Considering the approach proposed in this step, the time required for training and testing was reduced. A major step for the automatic description of images is the extraction of latent topic from the image scenery. Thus, in the next stage, latent topics based on the high level of extracted semantic regions were obtained by the help of bag of word and improved k-means methods. Considering the bag of visual word obtained from the region proposal, their locations and label of objectives were extracted using the supervised document neural autoregressive distribution estimator for the latent topics in the visual images. At the end, with the combination of these high-level features, classification of the image scenery was done. This proposed approach was applied to Scene-15, UIUC sports, MIT-67 datasets with 15, 8 and 67 scenes, respectively. The results obtained indicated that this method can improve the accuracy and reduce the time of the testing stage in these datasets.

Keywords: Image Scene Classification, Deep Neural Network, Content Extraction, Topic Modeling

Shahrood University of Technology

Faculty of Computer Engineering

PhD Thesis in Artificial Intelligence Engineering

Content Based Image Classification By Topic Models

By:

Ali Ghanbari Sorkhi

Supervisor:

Prof. Hamid Hassanpour

Advisor:

Dr. Mansour Fateh

January 2019