مقاله پژوهشی

مجله ماشین‌بینایی و پردازش‌تصویر
سال نهم، شماره چهارم، زمستان ۱۴۰۱

شناسایی صحنه، مبتنی بر همجوشی در دادگان جدید چندطیفی (مرئی-فروسرخ) و شبکه‌های پیچشی ژرف، با رویکرد یادگیری انتقالی

رحمان سروش[۱]، یاسر بالغی[۲]

چکیده

در دهه‌های اخیر، تکنیک‌های مختلفی در حوزه بینایی کامپیوتر، برای طبقه‌بندی و شناسایی صحنه‌ها در فضاهای مختلف، بر روی تصاویر طیف مرئی ارائه شده است. در این مقاله، ابتدا یک پایگاه داده تصویری چند طیفی، شامل زوج تصاویر طیف مرئی رنگی و فروسرخ ایجاد می‌شود. سپس با تجزیه تصاویر طیف مرئی و فروسرخ، به وسیله تبدیل موجک و استفاده از یک روش وزن‌دهی مبتنی بر آموزش شبکه‌های عصبی پیچشی ژرف، همجوشی تصاویر انجام می‌شود. همچنین این رویکرد، با چندین روش همجوشی دیگر و با استفاده از معیارهای ارزیابی کمی، مقایسه می‌شود. در نهایت، با استفاده از معماری‌های مبتنی بر شبکه‌های عصبی پیچشی ژرف آموزش دیده، تصاویر صحنه‌های مختلف، طبقه‌بندی می‌شوند. برای آموزش این شبکه‌ها بر روی مجموع تصاویر این پایگاه داده کوچک، از رویکرد یادگیری انتقالی، استفاده می‌شود تا طبقه‌بندی صحنه، با کمترین هزینه محاسباتی انجام گیرد. نتایج تجربی نشان می‌دهند که روش پیشنهادی، در طبقه‌بندی صحنه، که به صورت همجوشی تصاویر چهارکاناله (RGB-IR) صورت گرفته است، کارآمد بوده و ضمن داشتن معیارهای کمی همجوشی بالاتر، منجر به عملکرد بهتر، در مقایسه با سایر رویکردهای همجوشی تصاویر چندطیفی و با دقت طبقه‌بندی ۹۶,۶۷% می‌شود.

کلیدواژه‌ها

شناسایی صحنه، تصاویر چندطیفی، همجوشی، شبکه‌های عصبی پیچشی ژرف، یادگیری انتقالی، تبدیل موجک

۱ مقدمه

صحنه، یک چشم‌انداز از محیط دنیای واقعی است که معمولا اشیاء و سطوح در آن، به شیوه‌ای معقول و منطقی، به گونه‌ای در کنار یکدیگر قرار می‌گیرند تا معنادار شوند.

در دهه۱۹۹۰، جذابیت‌هایی در زمینه تشخیص شی و صحنه، به وجود آمد. برخی از محققان، تشویق شدند تا برای تقسیم‌بندی مناظر شهری و تشخیص و جداسازی صحنه‌های داخلی از صحنه-های خارجی، روش‌هایی را ابداع کنند [۱]. هدف تشخیص صحنه[1]، طبقه‌بندی یک تصویر، به یک دسته‌بندی معنایی است که بهترین توصیف و خلاصه‌سازی محیط صحنه را داشته باشد. حوزه بینایی کامپیوتر، نقش مهمی را در تجزیه و تحلیل، شناسایی و اجرایی شدن این هدف، ایفا می‌کند.

تکنیک‌های اولیه شناسایی صحنه، قبل از اینکه در زمینه طبقه‌بندی کلی یک صحنه، مورد استفاده قرارگیرند، به شناسایی اشیای داخل صحنه می‌پرداختند. تکنیک‌هایی که برای شناسایی صحنه، استفاده

[1] Scene Recognition

این مقاله در آبان‌ماه سال ۱۴۰۰ دریافت، در دی‌ماه بازنگری و در بهمن‌ماه همان سال پذیرفته شد.

[۱] دانشجوی دکتری الکترونیک دیجیتال، دانشکده مهندسی برق و کامپیوتر، دانشگاه صنعتی نوشیروانی بابل، بابل، ایران
رایانامه: rahmansoroush50@gmail.com

[۲] دانشکده مهندسی برق و کامپیوتر، دانشگاه صنعتی نوشیروانی بابل، بابل، ایران
رایانامه: y.baleghi@nit.ac.ir

نویسنده مسئول: یاسر بالغی

DOR: dorl.net/dor/20.1001.1.23831197.1401.9.4.2.9

می‌شوند، به ویژگی‌های بصری بکاررفته در نمایش صحنه که بیانی حسی از درک صحنه است، بستگی دارند. رنگ، جزئیات بافت و مشتقات تصویر، از مهمترین پارامترهای شناسایی صحنه، در تکنیک‌های اولیه مبتنی بر استخراج ویژگی‌های سراسری[1] صحنه است. فرآیند پردازشی مغز و سیستم بینایی انسان، در درک تصاویر بصری، به گونه‌ای است که هر انسانی قادر است با یک نگاه اجمالی از صحنه، درک جامعی از محتوای صحنه داشته باشد. برخی دیگر از رویکردها و تکنیک‌ها، بر اساس یک روش تصمیم-گیری انسان، در درک محتوای صحنه، الهام گرفته‌اند. به عبارت دیگر، با توجه به وجود یک شئ منحصر بفرد و در یک مکان خاص، می‌تواند صحنه را طبقه‌بندی نماید. به عنوان نمونه، وجود یک هواپیما در یک مکان مشخص(مانند خیابان)، می تواند نشانه‌ای از صحنه فرودگاه باشد[۱].

استفاده از معماری‌های مختلف، بر مبنای شبکه‌های عصبی پیچشی ژرف[2]، یکی از رویکردهایی است که در موضوع طبقه-بندی، استفاده می‌شوند. به عنوان نمونه، می‌توان از معماری-های ResNet[۲]، GoogleNet[۳]، InceptionV3[۴] و غیره استفاده کرد. هر یک از معماری‌های فوق، بر روی مجموعه بسیار بزرگی از داده‌های تصویری مختلف، در جهت دستیابی به طبقه-بندی خاص خود، آموزش دیده‌اند. ضرایب، در لایه‌های پیچشی این معماری‌ها به خوبی و بر اساس ویژگی‌های تصویری، تنظیم شده‌اند و چگونگی استخراج این ویژگی‌ها را فرا گرفته‌اند. بنابراین، به جای آموزش ساختار معماری کامل شبکه، می‌توان لایه‌های انتهایی را حذف کرد. به جای لایه‌های انتهایی، می‌توان از طبقه-بندهای ساده و با تعداد نورون‌های کم که متناسب با مجموعه داده-های خاص باشند، به شکل یادگیری انتقالی[3] استفاده کرد. آموزش شبکه در این حالت، فقط در سطح ساده طبقه‌بند، صورت می‌گیرد و سایر لایه‌های پیچشی، غیر فعال[4] و در فرآیند آموزش، شرکت نمی‌کنند. این کار، نه تنها باعث کاهش بار محاسباتی، نسبت به استفاده از آموزش کامل یک شبکه پیچیده می‌گردد، بلکه به دلیل بهره‌گیری از یک شبکه از پیش‌آموزش دیده شده، مشکل نیاز به داده‌های آموزشی بالا را مرتفع می‌نماید [۵-۷].

کیفیت تصاویر طیف مرئی رنگی[5]، با پیشرفت تکنولوژی دستگاه-های تصویربرداری، به طور چشم‌گیری بهبود یافته است. اما شرایط محیطی، مانند کم بودن میزان روشنایی و شرایط بد آب و هوایی (مانند مه) می‌تواند روی آن تاثیر بگذارد. برای توانمندتر کردن دقت تصمیم‌گیری در طبقه‌بندی، می‌توان از اطلاعات و جزئیات تصاویر حاصل از حسگرهای مختلف تصویربرداری، استفاده بهینه کرد. از تکنیک‌های گسترده همجوشی[6] این تصاویر، استخراج ویژگی‌های مفید و همچنین از تجمیع اثر آنها می‌توان در جهت تسهیل فرآیند پردازش، استفاده کرد.

در این میان، همجوشی تصاویر فروسرخ[7] و طیف مرئی، در بسیاری از جنبه‌ها دارای برتری هستند. تصاویر طیف مرئی، از نور منعکس شده بدست می‌آیند، در حالی که تصاویر فروسرخ، از جذب اشعه حرارتی بدست می‌آیند. بنابراین این همجوشی، دارای اطلاعات بیشتری نسبت به سیگنال‌های تک طیفی است. همچنین تصاویر فروسرخ و طیف مرئی، تقریبا ویژگی‌های موجود در تمام اشیاء را بازنمایی می‌کنند. افزون بر این، این تصاویر را می‌توان با تجهیزات نسبتا ساده‌تری تهیه نمود. بنابراین، با همجوشی این تصاویر، می‌توان ضمن استفاده از اطلاعات موجود در رنگ و جزئیات بافت تصاویر طیف مرئی، از اطلاعات موجود در تصاویر فروسرخ نیز استفاده کرد [۸].

اکثر رویکردهای ارایه شده برای تشخیص صحنه، بر روی پایگاه داده‌هایی صورت گرفته است که تنها شامل تصاویر رنگی طیف مرئی هستند. در این مقاله، ابتدا یک پایگاه داده جدید تصویری چند طیفی، شامل زوج تصاویر طیف مرئی رنگی و فروسرخ که از یک صحنه، گرفته شده‌اند ایجاد شده است.

در رویکرد پیشنهادی، ابتدا برش‌هایی از تصاویر طیف مرئی و فروسرخ، برای آموزش به یک شبکه عصبی پیچشی داده می‌شوند تا در تعدادی از لایه‌های پیچشی و بر اساس میزان اطلاعات برجستگی پیکسل‌ها یک نقشه وزن‌دهی مبتنی بر زوج تصاویر طیف مرئی و فروسرخ را به دست دهد. سپس با تجزیه تصاویر ورودی به مولفه‌های فرکانسی مبتنی بر تبدیل موجک[8] و اعمال ضرایب وزنی، بر روی این هرم‌های تصویری، همجوشی مولفه‌ها صورت می‌گیرد. همچنین ضمن بررسی معیارهای ارزیابی کمی این نوع همجوشی با سایر روش‌ها، از معماری‌های مبتنی بر شبکه-های عصبی پیچشی ژرف با رویکرد یادگیری انتقالی نیز که بر روی مجموعه بسیار بزرگی از داده‌های تصویری مختلف، آموزش دیده-اند، برای طبقه‌بندی صحنه استفاده می‌گردد.

سهم اصلی کار، در این مقاله را می‌توان به شرح زیر، خلاصه کرد: ۱- ایجاد یک پایگاه داده تصویری چندطیفی، شامل زوج تصاویر طیف مرئی رنگی و فروسرخ، از صحنه‌های مختلف طبیعی ۲- ارائه یک روش همجوشی چهارکاناله، از تصاویر طیف مرئی رنگی و فروسرخ، با استفاده از یک روش وزن‌دهی مبتنی بر آموزش شبکه‌های عصبی پیچشی (CNN)، برای دستیابی به یک تصویر با کیفیت و اطلاعات بصری بالاتر، بر اساس معیارهای کمی همجوشی ۳- استفاده از تکنیک یادگیری انتقالی، برای آموزش شبکه‌های پیچشی ژرف (DCNN) با مجموعه‌داده‌های آموزشی کوچک و کاهش هزینه محاسباتی در مقایسه با آموزش کامل شبکه

[1] Global
[2] Deep Convolutional Neural Networks (DCNN)
[3] Transfer learning
[4] Freeze
[5] Color visible (RGB)
[6] Fusion
[7] Infrared (IR)
[8] Wave LET

۴- افزایش دقت تشخیص در طبقه‌بندی صحنه نسبت به رویکردهای گزارش شده بر روی مسئله مشابه

بقیه مقاله، به شرح زیر سازماندهی می‌شود. در بخش ۲، کارهای مرتبط با زمینه همجوشی تصاویر طیف مرئی و فروسرخ، و شناسایی صحنه مبتنی بر شبکه‌های عصبی پیچشی ژرف، مورد مطالعه قرار می‌گیرند. در بخش ۳، مجموعه داده تصویری ایجادشده، ارائه می‌شوند. روش پیشنهادی، در بخش ۴ توضیح داده می‌شود. در بخش ۵، نتایج تجربی گزارش می‌شوند. نتیجه گیری در بخش ۶ ارائه می‌گردد.

۲ پژوهش‌های پیشین

بطورکلی، برای موضوع تشخیص مبتنی بر ترکیب تصاویر فروسرخ و طیف مرئی، از دو رویکرد همجوشی استفاده می‌شود. در رویکرد اول، ابتدا تصاویر را با هم ترکیب کرده و سپس تصاویر حاصل را به الگوریتم‌های تشخیص، اعمال می‌کنند. در رویکرد دوم، در حین اجرای الگوریتم‌های تشخیص، فرآیند همجوشی تصاویر انجام می‌شود. رویکرد دوم، باعث می‌شود که تعیین مرزبندی این دو فرآیند، دشوار گردد و معمولا استفاده از رویکرد اول متداول‌تر است. الگوریتم‌هایی که بر اساس رویکرد اول، عمل می‌کنند برای دستیابی به دید بهتر، غالبا در روش تجزیه تصاویر و یا قوانینی که برای همجوشی تصاویر، وضع می‌کنند با هم تفاوت دارند.

۲-۱ تکنیک‌های همجوشی تصاویر فروسرخ و قابل مشاهده

تصاویری که از ترکیب و همجوشی تصاویر فروسرخ و قابل مشاهده حاصل می‌شوند، ضمن بهره‌مندی از برجستگی‌های بصری و مزایای دید حرارتی تصاویر فروسرخ، ویژگی‌های رنگ و جزئیات بافت تصاویر طیف مرئی را نیز همزمان با خود دارند. به عنوان نمونه، تشخیص چهره را می‌توان یکی از مهمترین کاربردهای شناخته شده‌ای دانست که از مزایای همجوشی تصاویر فروسرخ و قابل مشاهده استفاده می‌کند. بنابراین استفاده از این تصاویر می‌تواند به بهبود کارآیی تکنیک‌های تشخیص صحنه و شناسایی اشیا نیز کمک کند.

بطورکلی، تکنیک‌های پذیرفته شده در همجوشی تصویر را می‌توان به هفت رویکرد تقسیم کرد [۸]. این رویکردها عبارتند از تبدیل مقیاس چندگانه[1]، بازنمایی تُنُک[2]، حساس به نقاط دارای برجستگی بصری[3]، مبتنی بر زیرفضا[4]، شبکه‌های عصبی و رویکردهای هیبریدی.

در رویکردهای مبتنی برتبدیل مقیاس چندگانه، ابتدا تصاویر اصلی را با استفاده از تبدیل‌های متداولی مانند تبدیل کسینوسی گسسته[5]، موجک، هرم[6] تصاویر و نسخه‌های اصلاح شده آنها، به لایه-های مختلف، تجزیه می‌کنند. سپس لایه‌هایی را که تحت قوانین خاصی به هم مربوط هستند را با هم ترکیب می‌کنند. سرانجام با استفاده از تبدیل معکوس روش‌های به کار گرفته شده، بر روی لایه‌های تلفیقی، برای بازسازی تصاویر همجوشی شده، استفاده می‌کنند [۹-۱۱]. رویکردهای مبتنی بر بازنمایی تُنُک نیز برای بهبود کارآیی شان، از یادگیری یک فرهنگ لغت جامع، از روی تعداد زیادی از تصاویر طبیعی با کیفیت بالا استفاده می‌کنند [۱۲, ۱۳]. رویکرد مبتنی بر نقاط دارای برجستگی بصری نیز از نحوه عملکرد سیستم بینایی انسان، الهام می‌گیرد. این روش، بر این اساس استوار است که سیستم بینایی انسان، اغلب به اشیاء یا پیکسل‌هایی که نسبت به همسایگان‌شان از برجستگی‌های بصری بیشتری برخوردارند، توجه بیشتری دارد. این رویکردهای همجوشی، می‌توانند ضمن حفظ یکپارچگی نواحی برجسته شئ، کیفیت بصری تصاویر همجوشی شده را نیز بهبود بخشند [۱۴-۱۶].

هدف رویکرد مبتنی بر زیرفضا، قرار دادن تصاویر ورودی با ابعاد بالا، در فضاهای کم بُعد یا زیرفضاها است. در اکثر تصاویر طبیعی، اطلاعات مازادی وجود دارد بطوریکه با استفاده از فضای فرعی کم بُعد، مانند تجزیه به مولفه‌های اساسی[7]، می‌توان ضمن ثبت ساختار واقعی تصاویر اصلی، به کاهش میزان حافظه مصرفی و هزینه زمانی پردازش داده‌ها کمک کرد. هدف تجزیه به مولفه‌های اساسی، تبدیل متغیرهای احتمالی همبسته، به متغیرهای ناهمبسته-ای به نام مولفه‌های اساسی است، بطوریکه همزمان با حفظ اطلاعات داده‌های اصلی، ابعاد داده‌ها نیز کاهش یابد [۱۷, ۱۸]. شبکه‌های عصبی نیز از روی نمونه‌های آموزشی که برای یادگیری به آنها داده می‌شوند، خود را سازگار می‌کنند و به تنظیم وزن پارامترهای مدل شبکه می‌پردازند. استفاده از آنها می‌تواند از قوانین پیچیده همجوشی در روش‌های سنتی، دوری کند. این شبکه‌ها این مزیت را دارند که بدلیل الهام گرفتن از عملکرد مغز انسان و چگونگی تعامل آن با اطلاعات عصبی، دارای قابلیت عدم حساسیت به نویز نیز باشند [۱۹-۲۳]. اگرچه تکنیک‌های همجوشی تصاویر فروسرخ و قابل مشاهده، مزایا و معایبی دارند، اما مدل‌های هیبریدی می‌توانند با ترکیب مزایای حاصل از این رویکردها، به بهبود کیفیت تصویر کمک کنند [۱۴, ۲۴]. در مرجع [۲۵] یک معماری همجوشی تصاویر طیف مرئی و نزدیک فروسرخ، با استفاده از شبکه‌های عصبی پیچشی VGG16دو مرحله‌ای، به نام FusionNet پیشنهاد شده است. برای آموزش شبکه، ابتدا تصاویر طیف مرئی را به نویز گوسی چندمقیاسی،

[1] Multi-scale transform
[2] Sparse representation
[3] Visual saliency
[4] Subspace
[5] Discrete Cosine Transform (DCT)
[6] Pyramid
[7] Principal Component Analysis (PCA)

آلوده می‌کنند. سپس، برای حذف نویز شدید تصاویر طیف مرئی، از یک زیرشبکه، استفاده می‌شود.

۱-۱-۲ معیارهای ارزیابی میزان کارایی رویکردهای همجوشی تصاویر

برای ارزیابی عملکرد روش‌های مختلف همجوشی تصویر فروسرخ و طیف مرئی، روش‌های بسیاری ارائه شده است که می‌توانند برای مقایسه عملکرد رویکردهای مختلف همجوشی و به عنوان یک راهنما برای انتخاب این روش‌ها در برنامه‌های واقعی، مورد استفاده قرارگیرند. همچنین از روش‌های ارزیابی، می‌توان برای تنظیم پارامترهای رویکردهای همجوشی استفاده شود. این روش‌ها در چارچوب موضوعات مختلفی که مبتنی بر تئوری اطلاعات، تشابه ساختار، گرادیان تصویر، آمار و سیستم بینایی انسان است، قرار می‌گیرند. در این بخش، به طور خلاصه برخی پارامترهای مهم ارزیابی کیفیت تصویر همجوشی شده، معرفی می‌شوند [۸].

الف- انتروپی[1] (EN): میزان اطلاعات موجود در یک تصویر همجوشی شده را بر اساس تئوری اطلاعات اندازه‌گیری می‌کنند. هرچه انتروپی، بزرگتر باشد اطلاعات بیشتری در تصویر همجوشی شده، وجود دارد و عملکرد روش همجوشی، بهتر خواهد بود. با این حال، ممکن است انتروپی، تحت تأثیر نویز، قرارگیرد. هرچه نویز تصویر همجوشی شده، بیشتر باشد انتروپی، بزرگتر خواهد شد. بنابراین، معمولاً انتروپی، به عنوان یک ارزیابی کمکی، مورد استفاده قرار می‌گیرد.

ب- اطلاعات متقابل[2] (MI): یک شاخص کیفیت است که میزان اطلاعاتی را که از تصاویر منبع، به تصویر همجوشی شده، منتقل می‌شود اندازه‌گیری می‌کند. این شاخص، یک مفهوم اساسی در تئوری اطلاعات است و میزان وابستگی دو متغیر تصادفی را اندازه‌گیری می‌کند. یک مقدار بزرگ MI به این معنی است که اطلاعات چشم‌گیری از تصاویر منبع، به تصویر همجوشی شده منتقل می‌شود، که نشان دهنده عملکرد خوب همجوشی است.

پ- شاخص QAB/F: میزان اطلاعات لبه را که از تصاویر منبع، به تصویر همجوشی شده منتقل می‌شود اندازه‌گیری می‌کند و نشان می‌دهد که چه مقدار از اطلاعات لبه تصاویر منبع، در تصویر همجوشی شده حفظ شده است. QAB/F بزرگ به این معنی است که اطلاعات لبه چشم‌گیری، به تصویر همجوشی شده، منتقل می‌شود.

ت- انحراف استاندارد[3] (SD): یک مفهوم آماری است که توزیع و کنتراست تصویر همجوشی شده را منعکس می‌کند. نواحی که دارای کنتراست زیاد هستند به دلیل حساسیت سیستم بصری انسان، به کنتراست، همیشه توجه انسان را به خود جلب می‌کنند. از این رو، یک تصویر تلفیقی با کنتراست زیاد، معمولاً منجر به یک SD بزرگ می‌شود. به این معنی که تصویر همجوشی شده، به یک جلوه بصری خوب می‌رسد.

ث- فرکانس مکانی[4] (SF): فرکانس مکانی، یک شاخص کیفیت تصویر بر اساس گرادیان (یعنی شیب‌های افقی و عمودی) است که به ترتیب فرکانس ردیف و فرکانس ستون مکانی نیز گفته می‌شود. سنجش SF می‌تواند توزیع گرادیان یک تصویر را به طور مؤثر اندازه‌گیری کند و از این طریق، جزئیات و بافت یک تصویر را آشکار می‌کند. یک تصویر همجوشی شده با SF بزرگ، مطابق سیستم بصری انسان، نسبت به درک انسان، حساس است و دارای لبه‌ها و بافت‌های غنی است.

ج- شاخص کیفیت اطلاعات متقابل[5](FMI): برای اندازه‌گیری میزان اطلاعات ویژگی و بر اساس MI و اطلاعات ویژگی است که از تصاویر منبع، به تصویر همجوشی شده، منتقل می‌شود. یک مقدار بزرگ FMI به طورکلی، نشان می‌دهد که اطلاعات ویژگی قابل توجهی از تصاویر منبع به تصویر همجوشی شده منتقل می‌شود.

۲-۲ تکنیک های شناسایی صحنه

رویکردهای تشخیص صحنه را می‌توان به دو دسته کلی روش‌های سنتی و استفاده از شبکه‌های عصبی، تقسیم کرد. در این قسمت، برخی از رویکردهایی که از شبکه‌های عصبی، در زمینه شناسایی صحنه استفاده می‌شوند، مورد بررسی قرار می‌گیرند.

بطورکلی، برای تشخیص صحنه نیز از دو روش، استفاده می‌شود. یکی مبتنی بر ویژگی‌های سراسری و دیگری، مبتنی بر ویژگی‌های محلی[6] است. رویکردهای مبتنی بر ویژگی‌های سراسری، سعی می‌کنند تا با درنظرگرفتن کل تصویر صحنه، ویژگی‌های بصری سطح پایین[7] را از هر بخش یا مناطق جذاب‌تر درون صحنه، استخراج کنند. سپس از ترکیب آنها، یک هیستوگرام ویژگی می‌سازند. رویکردهای مبتنی بر ویژگی‌های محلی نیز با توجه به چگونگی توزیع ویژگی‌های سطح پایین آن، به مدل‌سازی مولفه‌های بصری (مانند کلمات بصری[8]، اشیاء و موضوعات معنایی) می‌پردازند. در صورتیکه اشیای صحنه، خیلی کوچک و یا به هم‌بسته باشند، استفاده از ویژگی‌های محلی، برای طبقه‌بندی صحنه، از دقت کافی برخوردار نیست. اگرچه استفاده از روابط معناشناختی و ویژگی‌های محلی آنها می‌تواند مفید باشند اما این رویکرد، به دلیل تنوع بسیار زیاد بین کلاسی که درصحنه های داخلی وجود دارند، در طبقه بندی صحنه‌های داخلی، کارآمد نیست.

---

[1] Entropy
[2] Mutual Information (MI)
[3] Standard Deviation (SD)
[4] Spatial Frequency (SF)
[5] Feature Mutual Information (FMI)
[6] Local
[7] Low level
[8] Visual words

در مرجع [۲۶-۲۸]، برای تشخیص صحنه، از یک رویکرد مبتنی بر واژگان اشیاء تصویر، به نام بانک شی[1] و همچنین درک و استخراج مفاهیم سطح بالا استفاده می‌کنند. در مرجع [۲۹]، یک شبکه عصبی پیچشی ژرف هیبریدی متشکل از معماری‌های [۳۰] AlexNet و [۳۱] VggNet، با مدل‌های مبتنی بر فرهنگ لغت را برای تشخیص صحنه، پیشنهاد کردند. در مرجع [۳۲]، نشان داده شد که استفاده از توصیف‌کننده‌های معنایی، برای حذف اثرات منفی و ویژگی‌های عمومی اشیایی که به طور مشترک در بین صحنه‌های مختلف، وجود دارند می‌تواند در بهبود دسته‌بندی صحنه، مفید باشند. آنها نشان دادند که بررسی احتمالاتی وقوع این اشیاء در صحنه‌های مختلف و یافتن بیشترین احتمال قرارگیری یک شئ خاص، در مکانی خاص، می‌تواند باعث ایجاد وجه تمایزی برای طبقه‌بندی صحنه‌های مختلف، با محوریت اشیای درون آن شود. در مرجع [۳۳]، یک شبکه عصبی پیچشی با اهداف چند منظوره[2]، پیشنهاد داده شده است که از ترکیب ویژگی های متنی[3] سراسری، برای بهبود تشخیص صحنه و شناسایی شی، شی، در یک معماری واحد استفاده می کند. در مرجع [۳۴]، از یک معماری شبکه عصبی، برای شناسایی اهداف و تشخیص صحنه‌های نظامی استفاده شده است. در این رویکرد، از اطلاعات و ترکیب روابط معنایی بین اهداف و صحنه‌های نظامی، در جهت بهینه سازی نتایج، استفاده شده است. در مرجع [۳۵]، یک مدل یادگیری انتقالی مبتنی بر ResNet را با استفاده از تلفیق ویژگی‌های چند سطح از لایه‌های انتخاب شده که دارای ویژگی‌های تصویری مختلف هستند، برای بهبود دقت طبقه‌بندی صحنه استفاده کردند.

در مرجع [۳۶]، با ترکیب سه معماری مستقل , AlexNet GoogLeNet و VggNet بعنوان یک معماری مجموعه‌ای[4]، برای تشخیص ناهنجاری‌های رفتاری در داده‌های ویدئویی که از صحنه‌های شلوغ گرفته می‌شود، استفاده شده است. در مرجع [۳۷]، توانستند صحنه‌های با پوشش زمین، ابر و سایه را در تصاویر ماهواره‌ای با وضوح بالا (زیر۱۰ متر)، با استفاده از یک معماری مجموعه‌ای، با دقت بالا طبقه‌بندی کنند. در مرجع [۳۸]، از یک معماری شبکه عصبی پیچشی ژرف، برای طبقه‌بندی رودخانه‌ها استفاده شده است. مجموعه داده ورودی، متشکل از تصاویر رنگی از ۱۱ رودخانه است که در کشورهای کانادا، ایتالیا، ژاپن، انگلستان و کاستاریکا گرفته شده است. در مرجع [۳۹]، از یک معماری واحد شبکه عصبی پیچشی ژرف با مجموعه‌ای از طبقه‌بندهای کمکی، برای طبقه‌بندی تصاویر صحنه، استفاده شده است. همزمان با استخراج ویژگی‌ها در لایه‌های مختلف شبکه، به هر ویژگی انتخاب شده در هر لایه، یک طبقه‌بند متناسب با آن، اعمال و آموزش داده می‌شود. با ترکیب وزنی این طبقه‌بندهای کمکی که در لایه‌های مختلف از ویژگی‌های تصویری، آموزش دیده‌اند، طبقه‌بندی نهایی صورت می‌گیرد.

در مرجع [۴۰]، مجموعه داده‌های تصویری با تمرکز چندگانه[5] را به یک معماری مجموعه‌ای، که از سه شبکه عصبی پیچشی ژرف تشکیل شده است، اعمال کردند. برای بدست آوردن یک نقشه اولیه، هر یک از شبکه‌ها با مجموعه داده متفاوت، آموزش داده می‌شوند. از نقشه بدست آمده مبتنی بر هر سه معماری، برای همجوشی تصاویر استفاده می‌شود. برخی از روش ها نیز برای افزایش کارایی ویژگی‌ها، با بکارگیری توصیف‌گرهای تبدیل ویژگی مستقل از مقیاس[6] چند طیفی [۴۱] بر روی کانال های R، G و B، از اطلاعات اضافی موجود در ویژگی‌های بافت و رنگ نیز استفاده می‌کنند. در مرجع [۴۲] یک معماری تلفیقی از شبکه پیچشی VGG19، برای همجوشی تصاویر طیف مرئی و فروسرخ، پیشنهاد شده است. در این کار، برای تشخیص اشیا و کشتی‌ها بر روی یک مجموعه‌داده دریایی واقعی، از اطلاعات تکمیلی استخراج شده از سه سطح پیکسل، ویژگی‌های لایه‌ی میانی و سطح طبقه‌بند شبکه پیچشی، استفاده شده است. سپس از یک زیرشبکه همجوشی و با کمک تصاویر نزدیک فروسرخ، برای بازیابی بافت‌های تصاویر طیف مرئی حذف شده، استفاده می-کنند. در مرجع [۴۳] دو مدل همجوشی تصویر، با استفاده از قطعه‌بندی معنایی تصاویر، در یک شبکه پیچشی ژرف، برای طبقه-بندی ۱۰ کلاس مختلف از یک صحنه، پیشنهاد شده است. در مرحله اولیه همجوشی، تصاویر طیف مرئی رنگی، به فضای رنگی CIELab، تبدیل می‌شوند. سپس، با استفاده از تبدیل موجک، مولفه درخشندگی طیف مرئی را با تصاویر قطعه‌بندی شده طیف نزدیک فروسرخ، ترکیب می‌کنند تا به یک تصویر با وضوح بالاتر، دست یابند. در مرجع [۴۴] برای شناسایی صحنه حاصل از تصاویر چند طیفی، از ترکیب دو شبکه متقارن مستقل و موازی از پیش آموزش‌دیده شده با معماری GoogleNet، که بر روی مجموعه داده‌های ImageNet [۴۵]، آموزش دیده‌اند استفاده شده است. برای آموزش، یکی از شبکه‌ها با تصاویر طیف مرئی، و دیگری با طیف نزدیک فروسرخ، تغذیه می‌شوند. سپس، با استفاده از ترکیب سه لایه‌ی متناظر، از پیش‌طبقه‌بندها در معماری دو زیرشبکه، شناسایی صحنه، صورت می‌گیرد. در مرجع [۴۶] یک طبقه‌بند تصاویر صحنه حاصل از طیف مرئی و نزدیک فروسرخ، با استفاده از یک شبکه عصبی پیچشی دو کاناله، پیشنهاد شده است. در این روش، ابتدا ویژگی‌های تصاویر، به صورت مستقل و به وسیله لایه‌های پیچشی اولیه هر زیرشبکه، استخراج می‌شوند. سپس، با ترکیب ویژگی‌ها در لایه کاملاً متصل[7]، طبقه‌بندی صحنه، انجام می‌گیرد.

[1] Object-bank
[2] Multi-task
[3] Context features
[4] Ensemble
[5] Multi- focus
[6] Scale Invariant Feature Transform (SIFT)
[7] Fully connected

۳ پایگاه داده

تصاویری که مؤلفه های آن، در یک زمان و در طول موج‌های طیفی مختلف، گرفته می شوند را تصاویر چندطیفی می‌نامند. دوربین‌های سنتی، بر پایه قوانین فیزیک نور و در طیف مرئی تابش امواج الکترومغناطیسی، که در محدوده ۰,٤ تا ۰,۷ میکرومتر قرار دارند، قابلیت تصویربرداری دارند. پیشرفت تکنولوژی و ساخت مواد حساس به سطوح انرژی در طول موج‌های مختلف، مانند نزدیک فروسرخ (محدوده ۰,۷۵ تا ۱,٤ میکرومتر)، فروسرخ کوتاه (محدوده ۱,٤ تا ۳ میکرومتر)، فروسرخ متوسط (محدوده ۳ تا ۸ میکرومتر)، فروسرخ بلند (محدوده ۸ تا ۱۵ میکرومتر) و فروسرخ دور (محدوده ۱۵ تا ۱۰۰۰ میکرومتر)، موجب شده است تا دریچه‌ای را به سوی امکانات و قابلیت‌های جدید در تصویربرداری، باز شود. برای ایجاد پایگاه داده‌ای که شامل زوج تصاویر طیف مرئی رنگی و فروسرخ باشد، از دو دوربین مجزای طیف مرئی و فروسرخ، مطابق شکل ۱ و با مشخصات جدول ۱ استفاده شده است.

شکل ۱ دوربین‌های طیف مرئی و فروسرخ مورد استفاده برای تولید دادگان، در این پژوهش

جدول ۱ مشخصات دوربین‌های طیف مرئی و فروسرخ

مشخصات	دوربین فروسرخ	دوربین طیف مرئی
	IR TECH T8	Canon IXUS 220 HS
Zoom	4X	5X
Spectral range	8-14μm	visible
Frame rate	50Hz	24Hz
CMOS module	3 million pixel	12 million pixel
LCD display	3.6″	2.7″
Format	JPG	JPG

ابتدا مطابق شکل ۲ و با سوار نمودن دوربین طیف مرئی، بر روی دوربین فروسرخ، تعدادی تصاویر آزمایشی، گرفته می‌شوند. پس از کالیبره کردن دو دوربین، با استفاده از این تصاویر، فرایند تصویربرداری از صحنه‌های مختلف طبیعی انجام می‌گیرد.

شکل ۲ نمایی از نصب دوربین‌ها روی یک پایه

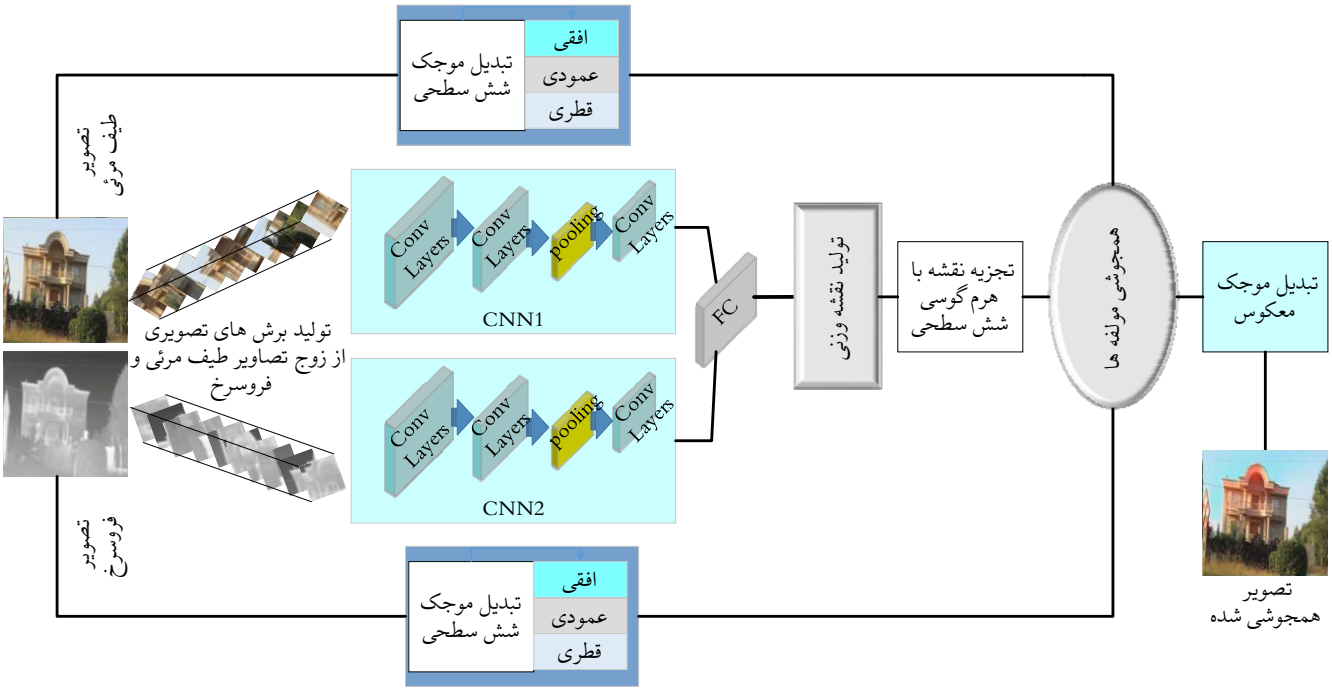
شکل ۳، نمونه‌هایی از زوج‌تصاویر طیف مرئی و فروسرخ صحنه-های مختلف را نشان می‌دهند. پایگاه داده حاصل، شامل ۱۰ دسته‌بندی است که مجموعا ۶۵۰ زوج تصویر از صحنه‌های مختلف را شامل می‌باشد. این صحنه‌ها شامل ساختمان، باغ مرکبات، مناطق ییلاقی، جنگل، کوه، مزرعه برنج، جاده، صخره، خیابان و مزرعه چای در مناطق مختلف شمال کشور ایران است. دسته‌بندی برخی از صحنه‌ها بسیار چالش‌برانگیز است. این به دلیل شباهت بسیار زیاد آنها در برخی از جزئیات صحنه (مانند مقایسه مناطق ییلاقی با کوه) است.

۴ روش کار

در این روش پیشنهادی که در شکل ۴ نشان داده شده است، برای به دست آوردن یک قانون وزن‌دهی همجوشی برای تعیین وزن در مقیاس‌های مختلف، از یک شبکه عصبی پیچشی استفاده می‌شود. این شبکه پیچشی دوقلو[1]، شامل دو زیرشبکه یکسان با معماری یکسان است که دارای پارامترها و وزن‌های مشابه می‌باشند. این شبکه‌ها معمولاً در کارهایی استفاده می‌شوند که شامل یافتن رابطه بین دو چیز قابل مقایسه و یا مقایسه نمونه‌های مشابه در مجموعه-های مختلف است. تنظیم پارامترها در هر دو زیرشبکه، برای یافتن شباهت ورودی‌ها، با مقایسه بردارهای ویژگی آنها انجام می‌شود. استفاده از معیارهای تشابه، در تشخیص چهره [۴۷] و تأیید امضا [۴۸] را می‌توان از جمله کاربردها در معماری آنها دانست. این شبکه‌ها به ویژه در مواردی که داده‌های کافی برای آموزش یک شبکه عصبی پیچشی ژرف و برای طبقه‌بندی تصاویر در تعداد کلاس‌های زیاد، وجود ندارند به خوبی عمل می‌کنند. زیرا با به اشتراک گذاشتن ضرایب وزنی بین زیرشبکه‌ها باعث می‌شوند تا برای یادگیری در مرحله آموزش، به پارامترهای کمتری نیاز داشته باشند و بنابراین می‌توانند نتایج خوبی را با مقدار نسبتاً کم داده-های آموزشی ایجاد کنند [۴۹].

[1] Siamese Network

شکل ۳ نمونه‌هایی از زوج‌تصاویر طیف مرئی و فروسرخ از ۱۰ کلاس صحنه مختلف



شکل ۴ نمایی از فرایند آموزش شبکه‌های عصبی پیچشی برای همجوشی تصاویر طیف مرئی و فروسرخ

هر شاخه شبکه دوقلو، دارای سه لایه پیچشی و یک لایه کاهنده ابعاد نقشه ویژگی، مبتنی بر انتخاب حداکثر مقدار[1] در یک پنجره مشخص است. اندازه پنجره و گام هر لایه پیچشی، به ترتیب ۳ × ۳ و ۱ تنظیم می‌شوند. اندازه پنجره و گام لایه کاهنده ابعاد نقشه ویژگی نیز به ترتیب ۲ × ۲ و ۲ تنظیم می‌گردند. سرانجام، نقشه ویژگی به دست آمده از خروجی هر شبکه، به هم می‌پیوندند تا به یک بردار ویژگی ۲۵۶ بعدی تبدیل شوند. سپس برای ایجاد یک امتیاز دو کلاسه، از یک تابع[2] که بتواند امتیازها را به یک توزیع احتمال بین مقادیر ۰ و ۱ تبدیل کند به طوری که بتوان آنها را به صورت احتمال، تفسیر کرد، استفاده می‌شود.

در مرحله آموزش، ابتدا با اعمال فیلتر چند مرحله‌ای گوسی با مقیاس‌های متعدد[3] و در ۵ سطح با تاری[4] متفاوت، بر روی تصاویر آموزشی ورودی، تصاویر جدیدی تولید می‌گردند. این تصاویر جدید آموزشی که برای شبیه‌سازی در تفاوت وضوح بین تصاویر ورودی، ایجاد می‌شوند، ضمن داده‌افزایی[5] در فرآیند آموزش شبکه دوقلو، باعث قوام مدل شده و در نتایج به دست آمده از این رویکرد، اثربخش هستند. سپس برش‌های تصویری بدون همپوشانی، از پیکسل‌های مکانی متناظر زوج تصاویر فروسرخ و طیف مرئی، با ابعاد ۱۶ × ۱۶ تولید می‌شوند. سرانجام، برای آموزش شبکه‌ها و بر اساس یک نمونه‌برداری تصادفی، تکه‌های آموزشی تصاویر مرئی، به شبکه CNN1 و تکه‌های تصاویر

[1] Max-Pooling
[2] Softmax function
[3] Standard deviations
[4] Blurry
[5] Data augmentation

فروسرخ متناظر، به شبکه CNN2 اعمال می‌شوند. مقدار احتمال هر کلاس، احتمال هر تخصیص وزن را نشان می‌دهد. با نرمالیزه کردن مجموع احتمالات دو خروجی شبکه‌ها، احتمال هر کلاس، دقیقاً وزنی را نشان می‌دهد که باید به طیف تصویر ورودی مرتبط، اختصاص داده شود. بدین ترتیب، یک نقشه وزنی با اندازه یکسان از تصاویر ورودی، به دست می‌آید.

در مرحله همجوشی، هر کدام از زوج تصاویر فروسرخ و طیف مرئی را با استفاده از تبدیل موجک، به مولفه‌های شش سطحی فرکانسی افقی، عمودی و قطری تجزیه می‌کنیم. همچنین ماتریس وزنی به دست آمده از شبکه عصبی نیز با استفاده از هرم گوسی[1]، به شش سطح تجزیه می‌شود تا بتواند در فرآیند همجوشی مولفه‌های فرکانسی تصاویر ورودی، بر روی هر دو طیف، اعمال شود.

در فرآیند همجوشی، ابتدا با استفاده از پنجره‌های کوچک ماتریسی، نقشه انرژی محلی محدود به هر پنجره در همه مولفه‌های تصویری، با استفاده از جمع مربعات ضرایب داخل هر پنجره، مطابق رابطه ۱ محاسبه می‌گردد.

$$[E(x,y)]^k_{IR\{H.D.V\}}=\sum_i\sum_j[WLT(x+i,y+j)^2]^k_{IR\{H.D.V\}}$$

$$[E(x,y)]^k_{Vis\{H.D.V\}}=\sum_i\sum_j[WLT(x+i,y+j)^2]^k_{Vis\{H.D.V\}} \quad (۱)$$

که در آن WLT تبدیل موجک، E انرژی، k تعداد سطوح هرم تصاویر و H، V و D نیز به ترتیب مولفه‌های افقی، عمودی و قطری تبدیل موجک هستند. همچنین برای تعیین چگونگی اعمال ضریب همجوشی، بر روی مولفه‌های فرکانسی طیف‌های زوج-تصاویر ورودی، از یک معیار شباهت سنجی مبتنی بر انرژی، مطابق رابطه ۲ که با Sim نشان داده شده است، استفاده می‌شود. این مقدار، در فاصله [۱, ۱-] قرار دارد و هرچه به عدد ۱ نزدیکتر باشد، میزان شباهت تصویر همجوشی شده، با طیف مربوط، بیشتر است [۲۰].

$$[Sim(x,y)]^k_{\{H,D,V\}}=\frac{2\sum_i\sum_j[WLT(x+i.y+j)^2]^k_{Vis\{H,D,V\}}\,[WLT(x+i.y+j)^2]^k_{IR\{H,D,V\}}}{[E(x,y)]^k_{Vis\{H,D,V\}}+[E(x,y)]^k_{IR\{H,D,V\}}} \quad (۲)$$

با درنظرگرفتن یک مقدار آستانه‌ای برای رابطه ۲، بر اساس تصاویر ورودی پایگاه داده، مولفه‌های فرکانسی تصویر همجوشی شده، مطابق روابط ۳ و ۴ به دست می‌آید.

$$[WLT(x,y)]^k_{Fused\{H,D,V\}}=\begin{cases}[GP(x,y)]^k_W\,[WLT(x,y)]^k_{IR\{H,D,V\}}\\ +(1-[GP(x,y)]^k_W)\,[WLT(x,y)]^k_{Vis\{H,D,V\}}\\ \text{اگر}\begin{cases}[Sim(x,y)]^k_{\{H.D.V\}}>\text{آستانه}\\ [E(x,y)]^k_{IR\{H,D,V\}}>[E(x,y)]^k_{Vis\{H,D,V\}}\end{cases}\end{cases} \quad (۳)$$

$$[WLT(x,y)]^k_{Fused\{H.D.V\}}=\begin{cases}[WLT(x,y)]^k_{IR\{H,D,V\}}\\ \text{اگر}\begin{cases}[Sim(x,y)]^k_{\{H,D,V\}}<\text{آستانه}\\ [E(x,y)]^k_{IR\{H,D,V\}}>[E(x,y)]^k_{Vis\{H,D,V\}}\end{cases}\\ \text{یا}\\ [WLT(x,y)]^k_{Vis\{H,D,V\}}\\ \text{اگر}\begin{cases}[Sim(x,y)]^k_{\{H,D,V\}}<\text{آستانه}\\ [E(x,y)]^k_{Vis\{H,D,V\}}>[E(x,y)]^k_{IR\{H,D,V\}}\end{cases}\end{cases} \quad (۴)$$

که در آن GP(x,y)، ماتریس وزنی تجزیه شده k سطحی با استفاده از هرم گوسی است. پس از همجوشی مولفه‌های تصویری، با استفاده از تبدیل معکوس موجک، مولفه‌های فرکانسی را با هم ترکیب کرده و تصویر همجوشی شده، بازسازی می‌شود. در نهایت، برای طبقه‌بندی تصاویر حاصل از همجوشی زوج تصاویر صحنه-های مختلف، از یک شبکه پیچشی ژرف با رویکرد یادگیری انتقالی استفاده می‌گردد.

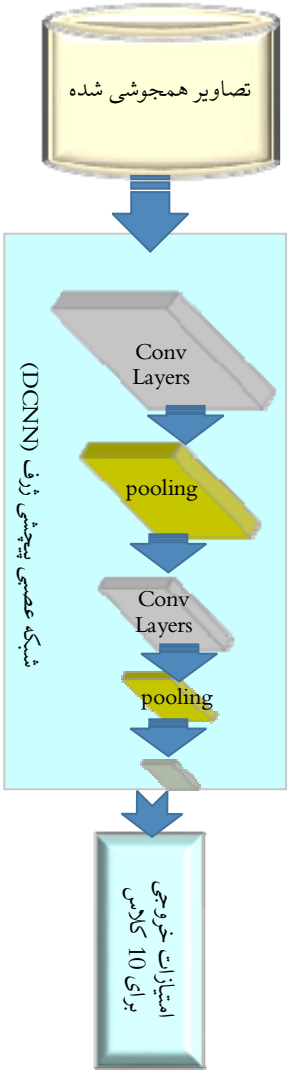
همچنین برای مقایسه اثر همجوشی روش پیشنهادی با سایر رویکردها در نتیجه طبقه‌بندی با یک شبکه پیچشی ژرف همسان، از روش‌های مختلف همجوشی مبتنی بر کانال‌های رنگی [۵۰]، نقاط دارای برجستگی بصری [۱۴] نیز استفاده می‌شود. با اعمال تصاویر حاصل از همجوشی‌های مختلف، به شبکه‌های عصبی پیچشی مستقل، با رویکرد یادگیری انتقالی و با امتیازدهی به نتایج خروجی طبقه‌بندها بر اساس تصاویر ورودی تست، ضمن محاسبه کمی معیارهای ارزیابی تصاویر همجوشی شده، ترکیب‌هایی که دارای بهترین مقادیر کمی می‌باشند و همچنین منجر به تصمیم-گیری بهتر در طبقه‌بندی صحنه، می‌گردند معرفی می‌شوند.

۱-۴ طبقه‌بندی صحنه، با تصاویر همجوشی شده

در طبقه‌بندی تصاویر صحنه‌های مختلف، که از روش‌های مختلف همجوشی به دست آمده‌اند، از شبکه های عصبی پیچشی مطابق شکل ۵ استفاده شده است. معماری شبکه‌های عصبی پیچشی، به گونه‌ای است که ویژگی‌های عمومی و کلی‌تر مجموعه داده‌های ورودی، مانند لبه‌ها را در لایه‌های اولیه، شکل‌ها را در لایه‌های میانی و برخی از ویژگی‌های خاص سطح بالا را در لایه-های بعدی، تشخیص می‌دهند. این شبکه‌ها عموما با میلیون‌ها

[1] Gaussian pyramid

تصویر ورودی از اشیای صحنه‌های مختلف، آموزش داده می‌شوند. آموزش کامل و اولیه این شبکه‌ها با میلیون‌ها پارامتر، به ویژه در مدل‌های پیچیده‌تر، زمان زیادی را برای آموزش، با استفاده از صدها دستگاه با پردازنده‌های گران‌قیمت، نیاز دارند. بنابراین برای کاهش هزینه و بار محاسباتی و جلوگیری از عدم کارآیی در طبقه‌بندی، به ویژه در مواردی که مجموعه داده‌ها کافی



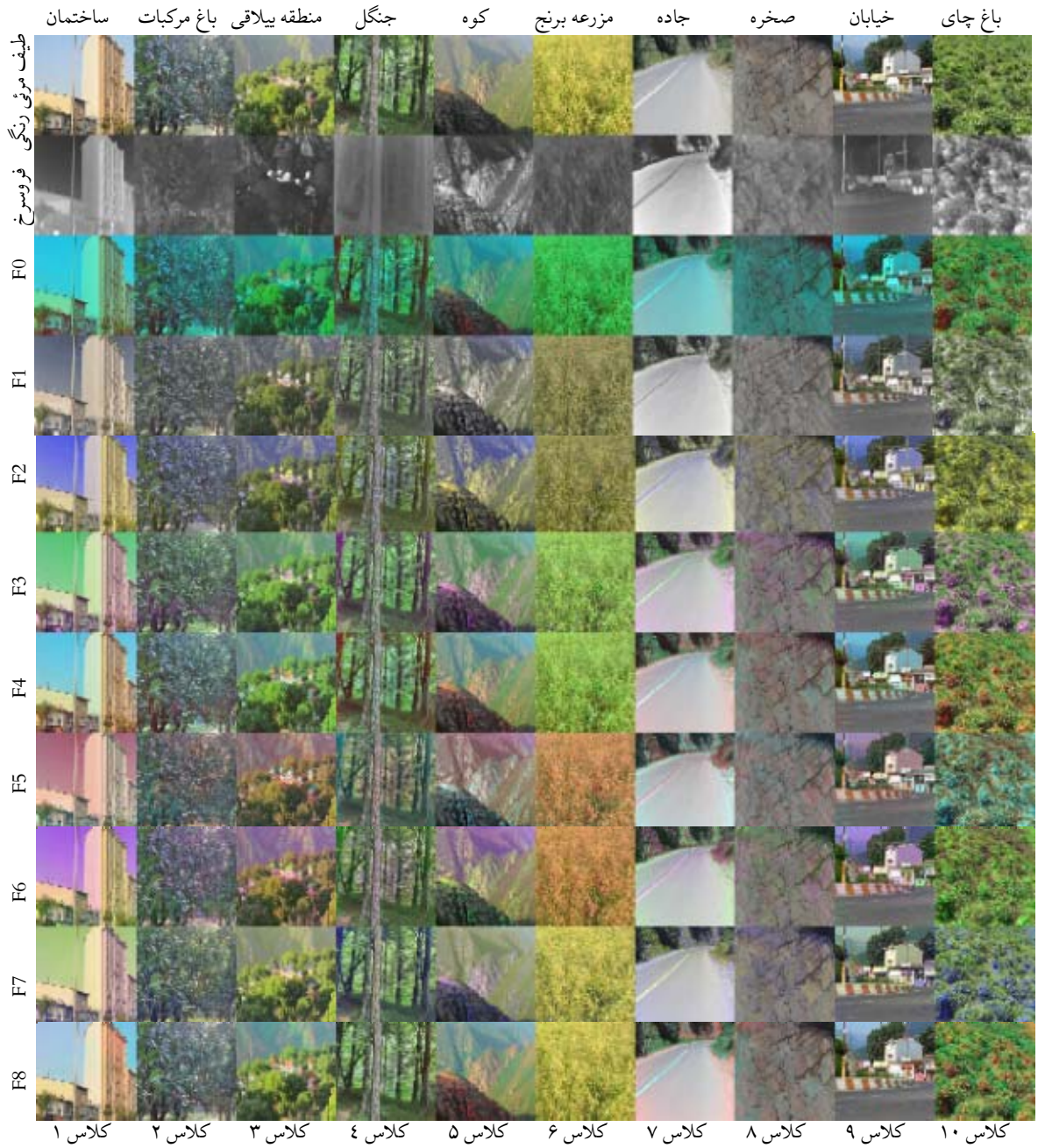
شکل ۵ نمایی از فرایند آموزش و طبقه‌بندی شبکه‌های عصبی پیچشی ژرف

نباشند، می‌توان از رویکرد یادگیری انتقالی، بر روی شبکه عصبی پیچشی آموزش‌داده شده، استفاده کرد. در این رویکرد، از اطلاعات و دانش حل یک مسئله، برای حل یک مسئله متفاوت اما مرتبط با آن استفاده می‌شود. بسته به بزرگی یا کوچکی مجموعه داده‌های جدید و میزان شباهت آنها با مجموعه داده‌های اولیه‌ای که شبکه عصبی پیچشی، به وسیله آن، آموزش دیده است، رویکردهای متفاوتی برای استفاده از یادگیری انتقالی، متصور است. اگر مجموعه داده‌های جدید، بزرگ و شبیه به مجموعه داده‌های اولیه باشند، می‌توان اطمینان بیشتری داشت که حتی اگر تمام شبکه، تحت آموزش قرارگیرد، باز هم دچار بیش‌برازش نمی‌شود. اگر مجموعه داده‌های جدید، کوچک و شبیه به مجموعه داده‌های اولیه باشند، اگرچه می‌توان انتظار داشت که ویژگی‌های سطح بالا در شبکه عصبی پیچشی نیز به این مجموعه داده، نزدیک باشند اما آموزش تمام شبکه، با مجموعه داده‌های ناکافی، عموما تمایل به بیش برازش مدل دارد. از این رو، ممکن است بهترین ایده، آموزش یک طبقه‌بند خطی، در انتهای شبکه عصبی پیچشی باشد. بنابراین، می‌توان آموزش تمام لایه‌ها را متوقف کرد و فقط طبقه‌بند نهایی را برای طبقه‌بندی خاصی که مورد نظراست، آموزش داد. همچنین درصورتی که مجموعه داده‌های جدید، کوچک و متفاوت از مجموعه داده‌های اولیه باشند، ممکن است لازم باشد تا ویژگی‌های موجود در چندلایه قبل نیز استخراج شوند و یک طبقه‌بند، در بالای آن ایجاد شود. در این حالت می‌توان چند لایه کاملا متصل را اضافه کرد و لایه‌های بالای شبکه عصبی را نیز آموزش داد.

اگرچه مجموعه داده‌های در دسترس این پایگاه داده، زیاد نیستند، اما بی‌شباهت با تصاویر آموزشی شبکه‌های عصبی پیچشی معروف مانند ResNet، GoogLeNet و غیره نیز نیستند. بنابراین، در این طبقه‌بندی نیز از رویکرد یادگیری انتقالی استفاده می‌شود. روش کار به این صورت است که ابتدا تعداد مناسبی از لایه‌های اولیه و میانی شبکه عصبی پیچشی از قبل آموزش دیده شده، در فرایند آموزش با مجموعه داده‌های جدید، شرکت داده نمی‌شوند تا وزن‌های این لایه‌ها دست نخورده باقی بمانند. سپس، سایر لایه‌های باقیمانده، با تصاویر خاص پایگاه داده ایجاد شده، آموزش داده می‌شوند تا وزن‌های این لایه‌ها با مجموعه داده جدید، سازگار شوند. جدول ۲ نام و فرایند همجوشی کانال‌های مختلف طیف مرئی با تصویر فروسرخ را نشان می‌دهد. در شکل ۶ نیز یک نمونه از تصاویر حاصل از این روش‌های همجوشی، در زوج‌تصاویر پایگاه داده، نشان داده شده است.

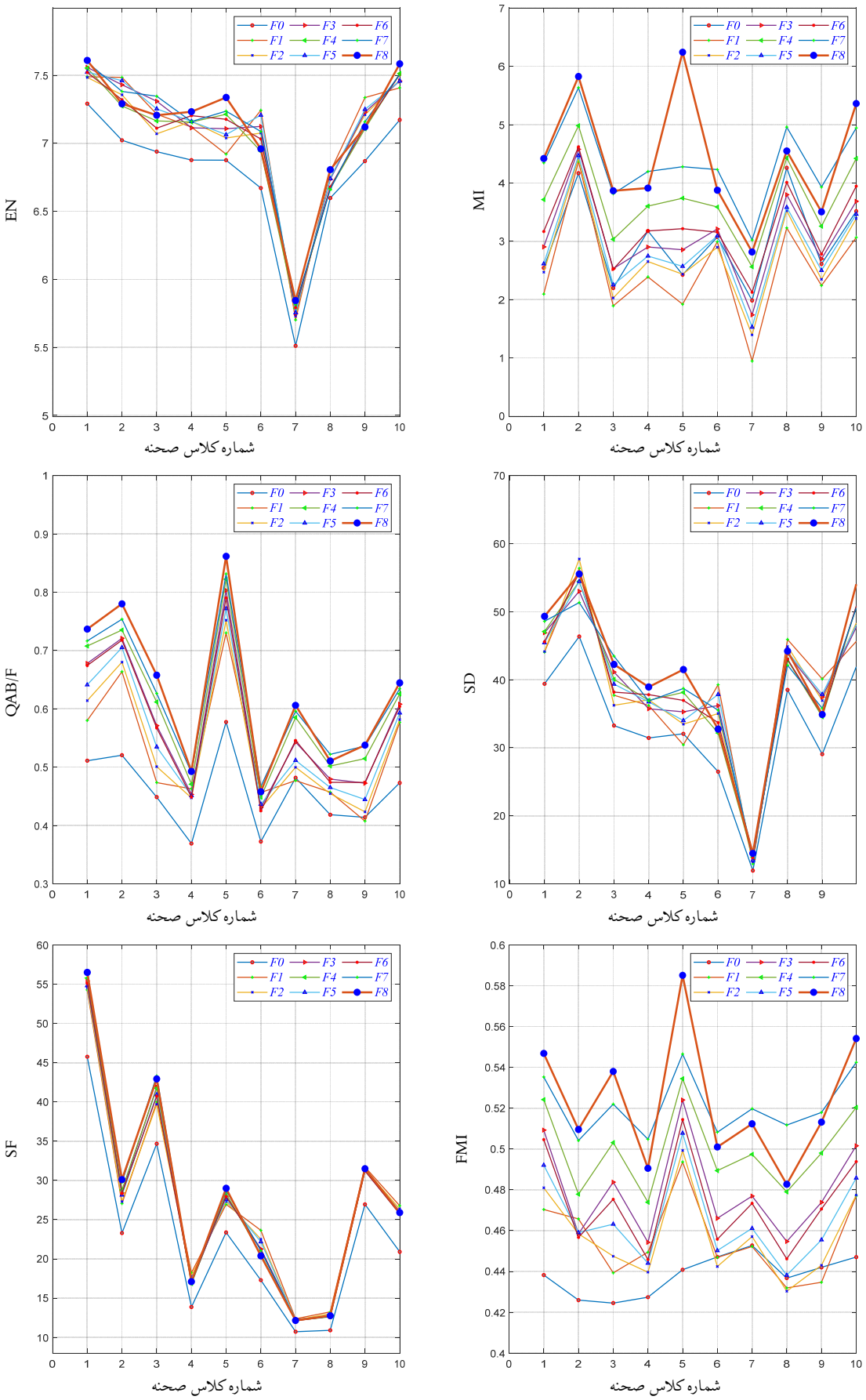
جدول ۲ نام و فرآیند همجوشی کانال‌های مختلف

نام‌گذاری	همجوشی	روش
F0	CH2-IR	CBCF [۵۰]
F1	CH1-IR CH2-IR CH3-IR	همجوشی کانال‌های طیف مرئی و فروسرخ مبتنی بر نقاط دارای برجستگی بصری [۱۴]
F2	CH1-IR CH2-IR	
F3	CH1-IR CH3-IR	
F4	CH1-IR	
F5	CH2-IR CH3-IR	
F6	CH2-IR	
F7	CH3-IR	
F8	تبدیل موجک مبتنی بر CNN	روش پیشنهادی



شکل ۶ یک نمونه از تصاویر حاصل از همجوشی‌های مختلف تصاویر ورودی RGB و IR

در بخش اول این کار، زوج‌تصاویر طیف مرئی و فروسرخ، با استفاده از روش‌های مختلف همجوشی، با هم ترکیب می‌شوند. تصاویر همجوشی شده باید ویژگی‌های طیف مرئی و فروسرخ در تصاویر را دارا باشند. بنابراین میزان تطابق تصاویر همجوشی شده، با تصاویر ورودی را می‌توان با استفاده از معیارهای کمی، ارزیابی کرد. برای این منظور، تعداد ۱۰ زوج تصویر از ۱۰ کلاس، از صحنه‌های مختلف که مجموعا شامل ۲۰۰ تصویر است، درنظر گرفته می‌شوند. مقایسه ۶ معیار ارزیابی EN, MI, QAB/F, SD, SF, FMI بر روی تصاویر حاصل از همجوشی‌های مختلف که بر روی ۱۰ کلاس تصویری، انجام گرفته است، در شکل ۷ نشان داده شده است. یک مقدار بزرگ برای هر یک از معیارهای کمی ارزیابی، به این معنی است که اطلاعات چشمگیری از تصاویر منبع (طیف مرئی و فروسرخ)، به تصویر همجوشی شده منتقل می‌شود که نشان دهنده عملکرد خوب همجوشی است.



شکل ۷ مقایسه معیارهای ارزیابی همجوشی‌های مختلف، بر روی ۱۰ کلاس تصویری

جدول ۳، درصد سهم اختصاص یافته به هر یک از روش‌های همجوشی را بر اساس بالاترین میزان هر یک از معیارهای ارزیابی در ۱۰ کلاس صحنه، گزارش می‌دهد. این نتایج، نشان می‌دهند که همجوشی پیشنهادی (F8)، در مقایسه با سایر روش‌های همجوشی، در اکثر کلاس‌ها بیشترین مقادیر معیارهای کمی ارزیابی همجوشی را دارد.

جدول ۳ درصد سهم اختصاص یافته به هر یک از روش‌های همجوشی، در ۱۰ کلاس صحنه

نام همجوشی	EN	MI	QAB/F	SD	SF	FMI
F0	۰	۰	۰	۰	۰	۰
F1	۳۰	۰	۰	۳۰	۶۰	۰
F2	۰	۰	۰	۱۰	۰	۰
F3	۰	۰	۰	۰	۰	۰
F4	۰	۰	۰	۰	۰	۰
F5	۰	۰	۰	۰	۰	۰
F6	۰	۰	۰	۰	۰	۰
F7	۱۰	۵۰	۲۰	۱۰	۱۰	۵۰
F8	۶۰	۵۰	۸۰	۵۰	۳۰	۵۰

در بخش دوم و برای دسته‌بندی صحنه‌های مختلف، ابتدا تصاویر حاصل از این همجوشی‌ها به شبکه‌های عصبی پیچشی از پیش آموزش دیده شده و با رویکرد یادگیری انتقالی اعمال می‌گردند. در این کار، از معماری ResNet50 استفاده شده است. این معماری، یک شبکه عصبی پیچشی ژرف ۵۰ لایه، با اندازه ورودی تصویر ۲۲۴ × ۲۲۴ است که از قبل، بر روی بیش از یک میلیون تصویر از پایگاه داده ImageNet آموزش دیده است. بنابراین، شبکه‌ای است که ویژگی‌های غنی را از طیف وسیعی از تصاویر، آموخته است و در مقایسه با سایر معماری‌ها، ضمن داشتن تعداد لایه‌های پیچشی کم که برای آموزش آن می‌توان از داده‌های کمتری استفاده کرد، خطای طبقه‌بندی پایین‌تری نیز دارد [۲].

شبیه‌سازی‌ها، با استفاده از برنامه MATLAB 2019b، بر روی یک سخت‌افزار با مشخصات جدول ۴، اجرا شده است.

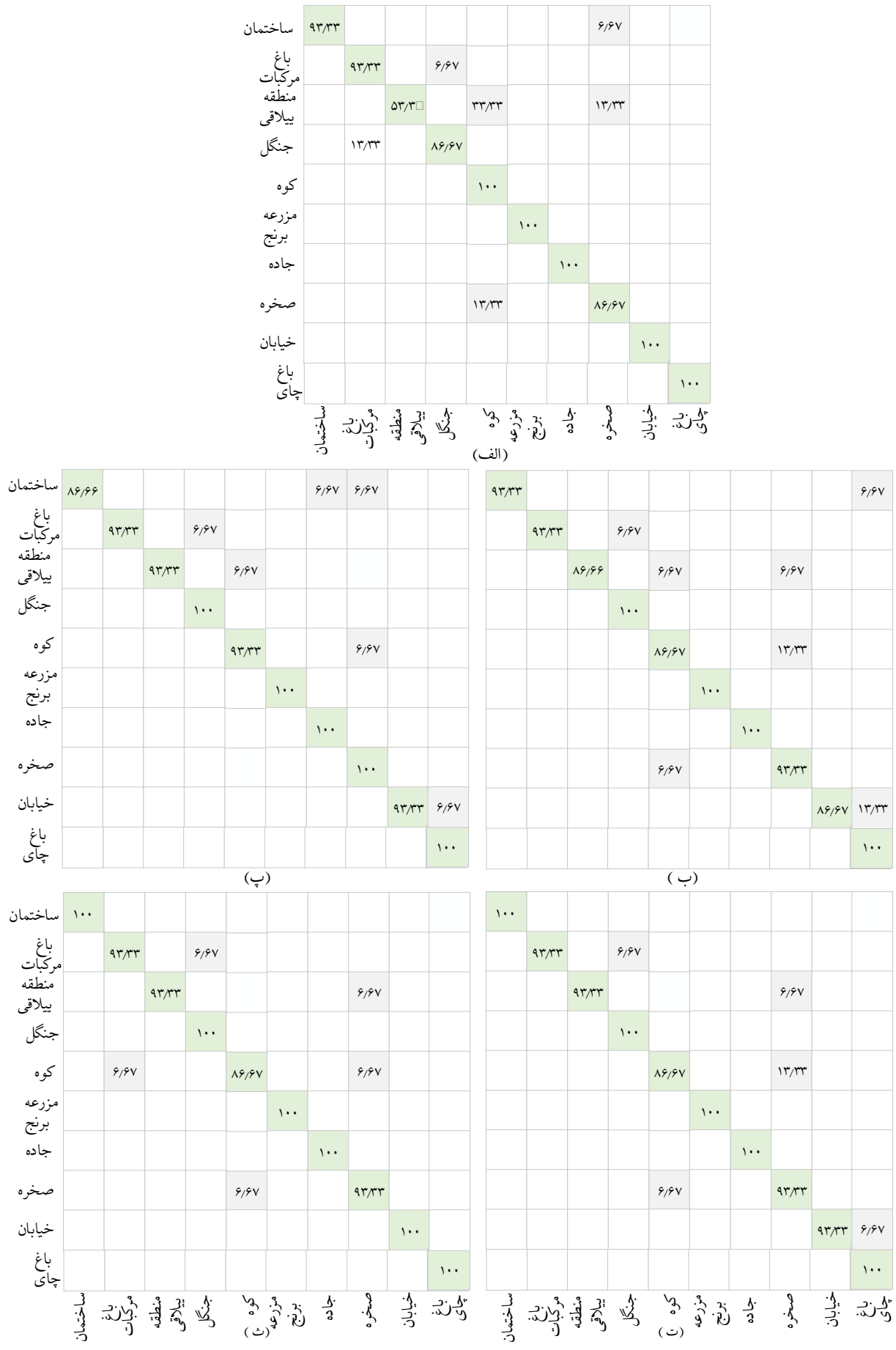
جدول ۴ مشخصات سخت‌افزار اجرای برنامه شبیه‌سازی

سخت‌افزار	مشخصات
System Type	ACPI x64-based PC
Processor	Intel(R) Core (TM) i7-6700 CPU @ 4.00GHz
GPU	NVIDIA GeForce GTX 980 Ti
RAM	12.0 GB

در مرحله اول و به صورت تصادفی، ۷۰ درصد از تصاویر آموزشی ورودی، برای آموزش شبکه و۳۰ درصد نیز برای ارزیابی ، انتخاب می‌شوند. تنها ۱۰ لایه‌ی انتهایی شبکه، در فرآیند آموزش، شرکت داده شدند و پارامترهای ۴۰ لایه‌ی اولیه آن، دست‌نخورده باقی ماندند.

پس از آموزش شبکه، تصاویر آزمونی که قبلا توسط شبکه دیده نشده است، به آن داده می‌شود تا به کلاس‌های مختلف صحنه، طبقه‌بندی نماید. سپس نتایج بدست آمده از طبقه‌بندی صحنه‌ها که با تصاویری از همجوشی روش‌های مختلف، حاصل شده است با یکدیگر مقایسه می‌شوند. برای یکسان‌سازی تعداد تصاویر در هر صحنه، تعداد ۶۵ تصویر از هر کلاس، انتخاب می‌گردند. تعداد ۵۰ تصویر از هر کلاس، برای آموزش شبکه و تعداد ۱۵ تصویر از هر کلاس نیز برای مرحله تست، استفاده می‌شوند. شکل ۸، جدول هم‌رخدادی[۳۹] دقت طبقه‌بندی تصاویر تک طیف مرئی و تصاویر حاصل از چهار روش همجوشی شده را که دارای معیارهای ارزیابی بالاتری هستند، بر اساس آموزش شبکه عصبی پیچشی ResNet50 با رویکرد یادگیری انتقالی، بر روی هر کلاس از تصاویر پایگاه داده نشان می‌دهد. میانگین دقت طبقه-بندی تصاویر تک طیف مرئی رنگی اصلی، ۹۱,۳۳ درصد است و بیشترین خطای آن، در تشخیص تصاویر دو صحنه چالش برانگیز، یعنی کوه و منطقه ییلاقی است که شباهت بسیار زیادی با هم دارند. میانگین دقت روش پیشنهادی، در طبقه‌بندی تصاویر همجوشی شده، با بهبود بیش از ۵ درصد، به ۹۶,۶۷ درصد افزایش یافته است.

[39] Confusion Matrix



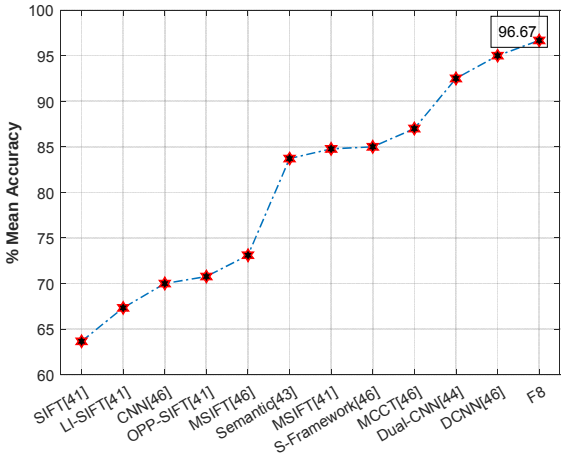
شکل ۸ دقت طبقه‌بندی تصاویر طیف مرئی رنگی (الف) و تصاویر حاصل از روش‌های همجوشی (ب-F1، پ-F2، ت-F7، ث-F8)

زمان سپری شده و درصد میانگین دقت طبقه‌بندی تصاویر حاصل از روش‌های مختلف همجوشی طیف مرئی و فروسرخ، که با تعداد ۷۰۰ تکرار[1] در طی ۲۰ دوره[2] و با نرخ یادگیری[3] ۰٫۰۰۰۱ به دست آمده است، نیز در جدول۵ نشان داده شده است.

جدول ۵ زمان سپری شده و دقت طبقه‌بندی رویکردهای همجوشی

همجوشی	F0	F1	F2	F3	F4	F5	F6	F7	F8
زمان سپری شده	۱۳'۰۱"	۱۳'۰۱"	۱۳'۰۱"	۱۳'۰۲"	۱۴'۰۵"	۱۳'۰۱"	۱۳'۰۱"	۱۵'۰۵"	۱۳'۰۰"
میانگین دقت	%۹۵٫۳۳	%۹۴	%۹۶	%۹۵٫۳۳	%۹۴	%۹۴	%۹۵٫۳۳	%۹۶	%۹۶٫۶۷

در شکل ۹ نیز مقایسه روش پیشنهادی، با سایر رویکردهای گزارش شده بر روی مسئله مشابه، نشان داده شده است.



شکل ۹ مقایسه رویکردهای مختلف بر روی مسئله مشابه

نتایج تجربی به دست آمده از رویکرد تشخیص صحنه‌ها و مبتنی بر چندین رویکرد همجوشی پیشنهادی، نشان می‌دهند که اولا میزان دقت طبقه‌بندی تصاویر صحنه‌ها در تمام روش‌های مختلف همجوشی، از تصاویر تک طیف مرئی، بالاتر است. این نتیجه، تایید می‌کند که تصاویر حاصل از همجوشی، دارای اطلاعات بیشتری نسبت به حالت تک‌طیفی می‌باشند. ثانیا رویکرد همجوشی F8، در مقایسه با سایر همجوشی‌ها ضمن داشتن مقادیر بزرگتری از معیارهای ارزیابی کمی، در تصمیم‌گیری نهایی طبقه‌بندی تصاویر صحنه‌های مختلف با رویکرد شبکه‌های عصبی پیچشی، و همچنین سایر رویکردهای گزارش شده بر روی مسئله مشابه، از دقت بیشتری هم برخوردار است.

۶ جمع‌بندی

بیشترین کارهای گزارش شده پیشین در طبقه‌بندی تصاویر صحنه، بر روی تصاویر تک طیف مرئی رنگی انجام شده است. ما در این کار، یک مجموعه داده چالش برانگیز، ایجاد کرده‌ایم که بطور همزمان از یک صحنه طبیعی و از ۱۰ کلاس مختلف گرفته شده است و شامل زوج تصاویر طیف مرئی و فروسرخ است. این پایگاه داده تصویری، می‌تواند در پژوهش‌های مختلف پردازش تصویر، نیز مورد استفاده قرارگیرد. همچنین، تکنیک‌های مختلف همجوشی بین کانال‌های رنگی طیف مرئی با فروسرخ، بکارگرفته شده است تا بهترین نوع همجوشی چهارکاناله، بر اساس معیارهای کمی ارزیابی، انتخاب شود. ما ضمن بهره‌گیری از آموزش یک شبکه عصبی پیچشی، برای دستیابی به یک نقشه وزن‌دهی و با استفاده از تجزیه تصاویر به مولفه‌های فرکانسی تبدیل موجک، به یک همجوشی رنگی با معیارهای کمی بالاتر دست یافتیم. همچنین با استفاده از رویکرد تکنیک یادگیری انتقالی، بر روی شبکه عصبی پیچشی ژرف ResNet50، ضمن کاهش بار محاسباتی، این شبکه را بطور ویژه برای مجموعه داده‌ای کوچک، آموزش داده و دقت طبقه‌بندی صحنه را افزایش داده‌ایم.

۷ مراجع

[1] Farahzadeh, E., *Tools for visual scene recognition*. 2014, Nanyang Technological University.

[2] He, K., 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.

[3] Szegedy, C., W. Liu, Y. Jia, P. Sermanet, S. Reed, et al. *Going deeper with convolutions*. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015.

[4] Szegedy, C., V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. *Rethinking the inception architecture for computer vision*. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.

[5] Dietterich, T.G. *Ensemble methods in machine learning*. in *International workshop on multiple classifier systems*. 2000. Springer.

[6] Maji, D., A. Santara, P. Mitra, and D. Sheet, *Ensemble of deep convolutional neural networks for learning to detect retinal vessels in fundus images.* arXiv preprint arXiv:.04833 2016.

[7] Zhou, Z.-H., J. Wu, and W. Tang, *Ensembling neural networks: many could be better than all.* Artificial intelligence, 2002. **137**(1-2): p. 239-263.

[1] Iteration

[2] Epoch

[3] Learning rate

unit-linking PCNN in NSCT domain. Infrared Physics Technology, 2015. **69**: p. 53-61.

[23] Zhong, J., B. Yang, Y. Li, F. Zhong, and Z. Chen. *Image fusion and super-resolution with convolutional neural network*. in *Chinese Conference on Pattern Recognition*. 2016. Springer.

[24] Liu, Y., S. Liu, and Z.J.I.f. Wang, *A general framework for image fusion based on multi-scale transform and sparse representation*. 2015. **24**: p. 147-164.

[25] Jung, C., K. Zhou, and J. Feng, *FusionNet: Multispectral fusion of RGB and NIR images using two stage convolutional neural networks*. IEEE Access, 2020. **8**: p. 23912-23919.

[26] Li, L.-J., H. Su, F.-F. Li, and E. P Xing, *Object bank: A high-level image representation for scene classification & semantic feature sparsification*. 2010.

[27] Quattoni, A. and A. Torralba. *Recognizing indoor scenes*. in *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 2009. IEEE.

[28] Singh, S., A. Gupta, and A.A. Efros. *Unsupervised discovery of mid-level discriminative patches*. in *European Conference on Computer Vision*. 2012. Springer.

[29] Xie, G.-S., X.-Y. Zhang, S. Yan, and C.-L. Liu, *Hybrid CNN and dictionary-based models for scene recognition and domain adaptation*. IEEE Transactions on Circuits Systems for Video Technology, 2015. **27**(6): p. 1263-1274.

[30] Krizhevsky, A., I. Sutskever, and G.E. Hinton, *ImageNet classification with deep convolutional neural networks*. Communications of the ACM, 2017. **60**(6): p. 84-90.

[31] Simonyan, K. and A. Zisserman, *Very deep convolutional networks for large-scale image recognition*. arXiv preprint arXiv:.04833, 2014.

[32] Cheng, X., J. Lu, J. Feng, B. Yuan, and J. Zhou, *Scene recognition with objectness*. Pattern Recognition, 2018. **74**: p. 474-487.

[33] Sun, H., Z. Meng, P.Y. Tao, and M.H. Ang. *Scene recognition and object detection in a unified convolutional neural network on a mobile manipulator*. in *2018 IEEE international conference on robotics and automation (ICRA)*. 2018. IEEE.

[34] Chen, C., J. Huang, C. Pan, and X. Yuan. *Military image scene recognition based on CNN and semantic information*. in *2018 3rd International Conference on Mechanical, Control and Computer Engineering (ICMCCE)*. 2018. IEEE.

[35] Liu, S., G. Tian, and Y. Xu, *A novel scene classification model combining ResNet based transfer learning and data augmentation with a filter*. Neurocomputing, 2019. **338**: p. 191-206.

[36] Singh, K., S. Rajora, D.K. Vishwakarma, G. Tripathi, S. Kumar, et al., *Crowd anomaly detection using aggregation of ensembles of fine-tuned convnets*. Neurocomputing, 2020. **371**: p. 188-198.

[8] Ma, J., Y. Ma, and C. Li, *Infrared and visible image fusion methods and applications: A survey*. Information Fusion, 2019. **45**: p. 153-178.

[9] Li, S., B. Yang, and J. Hu, *Performance comparison of different multi-resolution transforms for image fusion*. Information Fusion, 2011. **12**(2): p. 74-84.

[10] Pajares, G. and J.M. De La Cruz, *A wavelet-based image fusion tutorial*. Pattern recognition, 2004. **37**(9): p. 1855-1872.

[11] Zhang, Z. and R.S.J.P.o.t.I. Blum, *A categorization of multiscale-decomposition-based image fusion schemes with a performance study for a digital camera application*. 1999. **87**(8): p. 1315-1326.

[12] Li, S., H. Yin, and L. Fang, *Group-sparse representation with dictionary learning for medical image denoising and fusion*. IEEE Transactions on biomedical engineering, 2012. **59**(12): p. 3450-3459.

[13] Wang, J., J. Peng, X. Feng, G. He, J. Fan, et al., *Fusion method for infrared and visible images by using non-negative sparse representation*. Infrared Physics, 2014. **67**: p. 477-489.

[14] Ma, J., Z. Zhou, B. Wang, and H. Zong, *Infrared and visible image fusion based on visual saliency map and weighted least square optimization*. Infrared Physics Technology, 2017. **82**: p. 8-17.

[15] Zhang, X., Y. Ma, F. Fan, Y. Zhang, and J. Huang, *Infrared and visible image fusion via saliency analysis and local edge-preserving multi-scale decomposition*. JOSA A, 2017. **34**(8): p. 1400-1410.

[16] Zhao, J., Y. Chen, H. Feng, Z. Xu, and Q. Li, *Infrared image enhancement through saliency feature analysis based on multi-scale decomposition*. Infrared Physics Technology, 2014. **62**: p. 86-93.

[17] Bavirisetti, D.P., G. Xiao, and G. Liu. *Multi-sensor image fusion based on fourth order partial differential equations*. in *2017 20th International conference on information fusion (Fusion)*. 2017. IEEE.

[18] Kong, W., Y. Lei, H. Zhao, and Technology, *Adaptive fusion method of visible light and infrared images based on non-subsampled shearlet transform and fast non-negative matrix factorization*. Infrared Physics Technology, 2014. **67**: p. 161-172.

[19] Kong, W., L. Zhang, and Y. Lei, *Novel fusion method for visible light and infrared images based on NSST—SF—PCNN*. Infrared Physics Technology, 2014. **65**: p. 103-112.

[20] Liu, Y., X. Chen, J. Cheng, H. Peng, and Z. Wang, *Infrared and visible image fusion with convolutional neural networks*. International Journal of Wavelets, Multiresolution Information Processing, 2018. **16**(03): p. 1850018.

[21] Liu, Y., X. Chen, H. Peng, and Z. Wang, *Multi-focus image fusion with a deep convolutional neural network*. Information Fusion, 2017. **36**: p. 191-207.

[22] Xiang, T., L. Yan, and R. Gao, *A fusion algorithm for infrared and visible images based on adaptive dual-channel*

[37] Shendryk, Y., Y. Rist, C. Ticehurst, and P. Thorburn, *Deep learning for multi-modal classification of cloud, shadow and land cover scenes in PlanetScope and Sentinel-2 imagery.* ISPRS Journal of photogrammetry remote sensing, 2019. **157**: p. 124-136.

[38] Carbonneau, P.E., S.J. Dugdale, T.P. Breckon, J.T. Dietrich, M.A. Fonstad, et al., *Adopting deep learning methods for airborne RGB fluvial scene classification.* Remote Sensing of Environment, 2020. **251**: p. 112107.

[39] Bai, S. and H. Tang, *Softly combining an ensemble of classifiers learned from a single convolutional neural network for scene categorization.* Applied Soft Computing, 2018. **67**: p. 183-196.

[40] Amin-Naji, M., A. Aghagolzadeh, and M. Ezoji, *Ensemble of CNN for multi-focus image fusion.* Information fusion, 2019. **51**: p. 201-214.

[41] Brown, M. and S. Süsstrunk. *Multi-spectral SIFT for scene category recognition*. in *CVPR* 2011. IEEE.

[42] Farahnakian, F. and J. Heikkonen, *Deep learning based multi-modal fusion architectures for maritime vessel detection.* Remote Sensing of Environment, 2020. **12**(16): p. 2509.

[43] Kumar, W.K., N.J. Singh, A.D. Singh, and K. Nongmeikapam, *Enhanced machine perception by a scalable fusion of RGB–NIR image pairs in diverse exposure environments.* Machine Vision Applications, 2021. **32**(4): p. 1-21.

[44] Ševo, I. and A. Avramović. *Multispectral scene recognition based on dual convolutional neural networks*. in *Proceedings of the 10th International Symposium on Image and Signal Processing and Analysis*. 2017. IEEE.

[45] Deng, J., W. Dong, R. Socher, L.-J. Li, K. Li, et al. *Imagenet: A large-scale hierarchical image database*. in *2009 IEEE conference on computer vision and pattern recognition*. 2009. Ieee.

[46] Jiang, J., F. Liu, Y. Xu, and H. Huang, *Multi-spectral RGB-NIR image classification using double-channel CNN.* IEEE Access, 2019. **7**: p. 20607-20613.

[47] Chopra, S., R. Hadsell, and Y. LeCun. *Learning a similarity metric discriminatively, with application to face verification*. in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*. 2005. IEEE.

[48] Bromley, J., J.W. Bentz, L. Bottou, I. Guyon, Y. LeCun, et al., *Signature verification using a "siamese" time delay neural network.* International Journal of Pattern Recognition Artificial Intelligence, 1993. **7**(04): p. 669-688.

[49] Zagoruyko, S. and N. Komodakis. *Learning to compare image patches via convolutional neural networks*. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015.

[50] Zheng, Y., *Image fusion and its applications*. 2011: BoD–Books on Demand.

رحمان سروش مقاطع تحصیلی کارشناسی مهندسی الکترونیک را در دانشگاه صنعتی خواجه نصیرالدین طوسی و کارشناسی ارشد را در دانشگاه تهران، گذرانده است. وی از سال ۱۳۹۵ دانشجوی دکتری الکترونیک دیجیتال دانشگاه صنعتی نوشیروانی بابل است. زمینه‌ی پژوهشی وی شامل پردازش تصویر و الکترونیک است.

یاسر بالغی مقاطع تحصیلی کارشناسی، کارشناسی ارشد و دکتری خود را در دانشگاه علم و صنعت ایران گذرانده است. وی از سال ۱۳۸۸ عضو هیات علمی دانشکده مهندسی برق و کامپیوتر دانشگاه صنعتی نوشیروانی بابل است. زمینه‌های پژوهشی وی شامل پردازش تصویر و الکترونیک دیجیتال است.