Journal of Applied And Basic Machine Intelligence Research (ABMIR)

پژوهش‌های نظری و کاربردی هوش ماشینی

سال سوم، شماره اول، بهار و تابستان ١٤٠٤، صفحات٥٠ - ٣٥

ارائه روشی نوین به منظور هم ترازی دادگان در الگوریتم‌های تلفیق تصاویر فروسرخ و RGB

راضیه رضوی[1]، رضا روحانی سروستانی[2]*

[1]کارشناسی ارشد، گروه مهندسی کامپیوتر، دانشکده فنی و مهندسی، دانشگاه شهرکرد، شهر شهرکرد، ایران

[2]استادیار، گروه مهندسی کامپیوتر، دانشکده فنی و مهندسی، دانشگاه شهرکرد، شهر شهرکرد، ایران

مقاله پژوهشی

تاریخ دریافت:
١٤٠٣/١١/٢٥

تاریخ پذیرش:
١٤٠٤/٤/٢٤

کلیدواژه‌ها:
تلفیق تصاویر، تشخیص حرکت انسان، هم‌تراز کردن

نویسنده‌مسئول:
rrohani@sku.ac.ir

چکیده

در سال‌های اخیر، تشخیص حرکت انسان به یکی از موضوعات مهم در حوزه بینایی ماشین تبدیل شده است. بااین‌حال، یکی از چالش‌های اساسی در این زمینه، استخراج ویژگی‌های مؤثر برای افزایش دقت تشخیص است. داده‌های ویدئویی فروسرخ و RGB معمولاً برای این منظور استفاده می‌شوند، اما هیچ‌کدام به‌تنهایی اطلاعات کاملی از صحنه ارائه نمی‌دهند. بنابراین، ترکیب این داده‌ها می‌تواند به استخراج ویژگی‌های دقیق‌تر منجر شود. یکی از راهکارهای مؤثر برای دستیابی به این هدف، استفاده از تکنیک‌های تلفیق اطلاعات است. بااین‌وجود، بیشتر مجموعه‌داده‌های تشخیص حرکت انسان برای تلفیق استانداردسازی نشده‌اند و داده‌ها به‌درستی با یکدیگر تراز نیستند. در این پژوهش، از مجموعه‌داده NTU RGB+D استفاده شده و با بهره‌گیری از تکنیک‌های مسائل معکوس و مختصات نقاط بدنی موجود در این مجموعه‌داده، روشی برای ترازسازی و برش داده‌های ویدئویی به‌منظور تلفیق دو نوع داده ویدئویی فروسرخ و RGB ارائه شده است. عملکرد روش پیشنهادی با استفاده از معیارهای EN، MI، SSIM و MS-SSIM مورد ارزیابی قرار گرفته است. نتایج به‌دست‌آمده نشان می‌دهند که مقادیر حاصله از EN(٧/١٧) و MI(١٣/١) بیانگر حداکثر میزان انتقال و هم‌پوشانی اطلاعات هستند. همچنین، مقادیر SSIM(٠/٧٨) و MS-SSIM (٠/٨٤) نشان‌دهنده حفظ ساختار و کیفیت بالای داده‌های تلفیق‌شده هستند. این نتایج، کارایی روش پیشنهادی را در بهبود تلفیق داده‌های ویدئویی تأیید می‌کنند.

doi: 10.22034/ABMIR.2025.22795.1101

E-ISSN: 2821-2037

۱- مقدمه

امروزه بیش از ۹۰ درصد دریافت اطلاعات از محیط پیرامون ما، به کمک مشاهده صورت می‌گیرد. با پیشرفت فناوری امکان ثبت، پردازش و انتقال تصاویر به صورت دیجیتال میسر شده است. پردازش تصویر به عنوان یکی از مولفه‌های اساسی در سیستم‌های هوشمند و پشتیبان تصمیم است که توسط سیستم‌های کامپیوتری بر روی تصاویر دیجیتال اعمال می‌شود. کاربردهای متنوعی که پردازش تصویر در زمینه‌های مختلف فنی، صنعتی، شهری، پزشکی و علمی دارد، آن را به یک موضوع بسیار فعال در میان زمینه‌های پژوهشی تبدیل کرده است. به دلیل پیشرفت تکنولوژی و نیاز به کسب اطلاعات فشرده از چندین منبع، تلفیق تصاویر که روشی پرکاربرد در حوزه پردازش تصویر است به یک زمینه تحقیقاتی بسیار مهم تبدیل شده است[۱].

تلفیق به معنای ترکیب اطلاعات از دو یا چند تصویر با خصوصیات مختلف است. با استفاده از تکنیک تلفیق می‌توان تصویر جدیدی ایجاد نمود که هم‌زمان دارای خصوصیات بارز تصاویر منبع و اطلاعات بیشتری نسبت به تک‌تک تصاویر اصلی باشد[۱]. تلفیق تصاویر نه تنها باعث کاهش افزونگی و حجم اطلاعات می‌شود، بلکه خروجی به‌دست آمده برای انسان و یا اهداف ماشین نیز قابل درک‌تر است[۲]. با توجه به کاربرد و خروجی مورد نظر برای تلفیق تصاویر از روش‌های متنوعی از جمله PCA، ICA، Wavelet و... در سطوح مختلف تلفیق می‌توان استفاده نمود[۳]. عموماً تلفیق تصاویر در سه سطح یعنی سطح پیکسل[1]، سطح ویژگی[2] و سطح تصمیم[3] صورت می‌گیرد[٤]؛ که استفاده از هر یک از سطوح ذکرشده به داده‌ها و کاربرد مورد نظر وابسته است[٥][٦].

یکی از کاربردهای مهم دیگری که امروزه در زمینه تلفیق اطلاعات مورد توجه قرار گرفته است، تشخیص حرکت انسان است[4]. در این حوزه می‌توان اطلاعات مختلف ثبت شده از حرکت بدن افراد را تلفیق نموده و اطلاعات بسیار مفیدی از حالت قرارگیری افراد به دست آورد، که این داده‌ها برای اهدافی همچون رباتیک، ساخت بازی‌ها، سیستم‌های رانندگی خودکار و... مورداستفاده قرار می‌گیرند [۷].

در حوزه تشخیص حرکت انسان تصاویری با مشخصه‌های مختلف مانند تصاویر RGB، فروسرخ و... می‌توانند در تلفیق به کار گرفته شوند [۸]. با توجه به آنکه استفاده از تصاویر RGB و یا تصاویر فروسرخ به تنهایی در مدل به دلیل وابستگی بالا به منابع نوری در تصاویر رنگی و یا حذف ناخواسته بعضی از ویژگی‌های تصویر توسط تصاویر فروسرخ محدودیت‌هایی به دنبال دارد، لذا می‌توان برای ساخت تصویری با ویژگی‌های قوی‌تر، از روش‌های تلفیق بر روی دو نوع تصویر RGB و تصاویر فروسرخ استفاده نمود[۹].

تلفیق داده‌های ویدئویی فروسرخ و RGB می‌تواند رویکردی مؤثر برای ارائه ویژگی‌های قوی‌تر به‌عنوان ورودی مدل‌های تشخیص حرکت انسان باشد. تأمین ویژگی‌های قدرتمند، عاملی کلیدی در افزایش دقت سامانه‌های تشخیص حرکت محسوب می‌شود. بنابراین، استفاده از داده‌های تلفیقی می‌تواند جایگزینی مناسب برای استفاده جداگانه از هر نوع داده در مدل‌های تشخیص حرکت انسان باشد.

یکی از شروط اصلی برای تلفیق تصاویر این است که تصاویر منبع باید کاملاً تراز باشند تا بتوان آن‌ها را ترکیب نمود. تراز بودن داده‌های ویدیویی به معنای هم‌راستایی دقیق اطلاعات موجود در دو مجموعه داده است، به‌طوری که ویژگی‌ها و اجزای مشابه در هر دو ویدئو در موقعیت‌های زمانی و مکانی یکسان قرار بگیرند. این فرایند اطمینان می‌دهد که داده‌ها برای ترکیب و تحلیل‌های بعدی هماهنگ و قابل‌مقایسه هستند. با این حال، در حوزه تشخیص حرکت انسان، هیچ مجموعه داده‌ای داده‌های ویدئویی فروسرخ و RGB تراز شده را ارائه نمی‌دهد، بنابراین لازم است قبل از انجام عملیات تلفیق بر روی داده‌ها آن‌ها را با یکدیگر تراز کرد.

در سال‌های گذشته، مجموعه‌داده‌های متعددی به‌منظور پیاده‌سازی مدل‌های مختلف و ارزیابی سیستم‌های تشخیص حرکت انسان

---

1 Pixel-Level
2 Feature-Level
3 Decision- Level
4 Human Action Recognition

توسط محققان جمع‌آوری و ارائه شده‌اند. یکی از شناخته‌شده‌ترین مجموعه‌داده‌ها در این زمینه NTU RGB+D است که شامل داده‌های ویدئویی فروسرخ و RGB است [۱۰]. بااین‌حال، این مجموعه‌داده به‌طور خاص برای تلفیق تصاویر استانداردسازی نشده است و داده‌های ویدئویی آن با یکدیگر تراز نیستند. درواقع، اندازه دو نوع داده RGB و فروسرخ در این مجموعه یکسان نبوده و به هم‌راستایی نیاز دارند.

در این پژوهش، با بهره‌گیری از روش‌های مسائل معکوس و مختصات نقاط بدنی موجود در مجموعه‌داده، روشی برای ترازسازی و برش فریم‌های ویدئویی فروسرخ و RGB پیشنهاد شده است. هدف این مطالعه، تولید یک مجموعه‌داده جدید از طریق تکنیک‌های مختلف تلفیق اطلاعات است، به‌طوری که بتوان از آن به‌عنوان ورودی برای سیستم‌های تشخیص حرکت انسان استفاده کرد. بهره‌گیری از داده‌های ویدئویی تلفیق‌شده به‌عنوان ورودی، ویژگی‌های قوی‌تری را در اختیار مدل قرار داده و درنهایت، دقت تشخیص را بهبود می‌بخشد.

در بخش بعدی، به بررسی مطالعات انجام‌شده در زمینه تلفیق تصاویر با ویژگی‌های مختلف پرداخته خواهد شد. در بخش سوم، مروری بر مسائل معکوس خواهیم داشت و سپس در بخش چهارم، مجموعه‌داده NTU RGB+D که در این پژوهش مورد استفاده قرار گرفته است، معرفی می‌شود. در بخش پنجم، روش پیشنهادی به‌صورت کامل ارائه خواهد شد و در نهایت، در بخش ششم، نتایج به‌دست‌آمده بررسی و تحلیل می‌شوند.

۲- کارهای پیشین

نیاز به تلفیق داده‌ها در سیستم‌های پردازش تصویر به دلیل پیشرفت تکنولوژی و نیاز به کسب اطلاعات فشرده از چندین منبع، روز به روز در حال افزایش است. تلفیق تصاویر نه‌تنها باعث کاهش حجم اطلاعات می‌شود بلکه خروجی به دست آمده برای انسان و یا اهداف ماشین نیز قابل درک‌تر خواهد بود. این تکنیک به‌عنوان یک روش پرکاربرد در زمینه‌های مختلفی همچون علم پزشکی، بینایی ماشین، انواع نظارت بر محیط و... به‌کار گرفته می‌شود.

در حوزه‌ی تلفیق تصاویر روش‌های سنتی همچون PCA [۱۱]، Hierarchical PCA [۱۲]، ICA [۱۱]، wavelet [۱۳] ، Pyramid [۱٤] و... در حوزه‌هایی همچون تلفیق تصاویر پزشکی مانند تصاویر PET و MRI ، تصاویر ماهواره‌ای، تلفیق در سطوح مختلف و بسیاری از موارد دیگر مورد استفاده قرار گرفته‌است.

اما با ظهور و رشد چشم‌گیر شبکه‌های عصبی در علوم کامپیوتر، بسیاری از پژوهشگران از این فناوری در جهت تلفیق و ادغام اطلاعات بهره گرفته‌اند و در نتیجه، به دستاوردهای قابل‌توجهی دست یافته‌اند. از جمله ساختارهای پرکاربرد در این حوزه، می‌توان به شبکه‌های عمیق کانولوشنی (CNN) اشاره کرد که به دلیل تولید خروجی‌های با کیفیت بالا، مورد توجه گسترده قرار گرفته‌اند [۱۵، ۱٦، ۱۷]. پژوهشگران با اتخاذ این شبکه‌ها به‌عنوان چارچوب پایه و اعمال اصلاحات ساختاری، توانسته‌اند عملکرد آن‌ها را در زمینه‌های مختلفی از جمله تلفیق تصاویر پزشکی و ترکیب تصاویر حاصل از سنسورهای متفاوت مانند تصاویر مادون‌قرمز و مرئی بهبود ببخشند[۱۸]. در ادامه، به بررسی برخی از مطالعات انجام‌شده در این زمینه خواهیم پرداخت.

در پژوهشی، لی و وو ساختاری مبتنی بر شبکه‌های عصبی کانولوشنی برای تلفیق تصاویر مادون‌قرمز و مرئی (Visible) پیشنهاد کرده‌اند [۱۹]. در روش پیشنهادی تصاویر ورودی که از پیش هم‌تراز شده‌اند، وارد شبکه‌ای با سه بخش اصلی شامل رمزگذار (Encoder)، لایه تلفیق (Fusion Layer) و رمزگشا (Decoder) می‌شوند. برخلاف شبکه‌های کانولوشنی متداول، در بخش رمزگذار از بلوک‌های متراکم (Dense Blocks) بهره گرفته‌شده است که در آن‌ها، خروجی هر لایه به تمامی لایه‌های بعدی متصل می‌شود. این ساختار موجب استخراج ویژگی‌های غنی‌تر و مؤثرتر از تصاویر ورودی می‌گردد. همچنین، در این پژوهش دو استراتژی تلفیق با استفاده از L1 norm و Addition برای ادغام ویژگی‌ها به‌کار رفته است. معماری پیشنهادی به‌گونه‌ای طراحی شده که با تصاویر ورودی در اندازه‌های مختلف سازگار بوده و از قابلیت تعمیم‌پذیری بالایی برخوردار است. ارزیابی‌های عینی و کیفی نشان داده‌اند که این روش نسبت به سایر روش‌های

موجود عملکرد برتری داشته و در بسیاری از معیارهای سنجش کیفیت، نتایج سطح بالایی را به‌دست آورده است.

ترکیب تصاویر مادون‌قرمز و مرئی با هدف تولید تصویری تلفیقی صورت می‌گیرد که بتواند مزایای هر دو منبع را به‌صورت هم‌زمان حفظ کند. در این راستا، شیائو و همکاران در پژوهشی، یک شبکه تلفیق نوین مبتنی بر Auto-Encoder پیشنهاد کرده‌اند. در این مدل، رمزگذار تصویر ورودی را به دو نگاشت ویژگی مجزا شامل نگاشت‌های پس‌زمینه و جزئیات تفکیک می‌کند که به‌ترتیب حاوی اطلاعات فرکانس پایین و بالا هستند، و رمزگشا وظیفه بازسازی تصویر نهایی را بر عهده دارد [۲۰]. همچنین، تابع هزینه به‌گونه‌ای طراحی شده است که شباهت در نگاشت‌های پس‌زمینه و تفاوت در نگاشت‌های جزئیات میان تصاویر ورودی به‌خوبی تقویت شود. در مرحله آزمون، این نگاشت‌ها توسط یک ماژول تلفیق ادغام شده و سپس تصویر نهایی از طریق رمزگشا بازسازی می‌شود. نتایج ارزیابی‌های کیفی و کمی نشان می‌دهد که روش پیشنهادی، ضمن استخراج ویژگی‌های قدرتمند، در حفظ بافت تصویر و برجسته‌سازی نواحی کلیدی عملکردی برتر نسبت به روش‌های پیشرفته موجود ارائه می‌دهد.

لی و همکاران روشی نوین برای تلفیق تصاویر مادون‌قرمز و مرئی ارائه داده‌اند که مبتنی بر شبکه‌های متصل تودرتو (Nested connection based) و مدل‌های توجه مکانی و کانالی است [۲۱]. ساختار پیشنهادی شامل سه بخش اصلی Encoder، استراتژی تلفیق و Decoder است. در این روش، ابتدا تصاویر ورودی به Encoder داده می‌شوند تا ویژگی‌های عمیق در سطوح مختلف استخراج شوند. سپس با بهره‌گیری از مکانیزم‌های توجه مکانی و کانالی، اهمیت هر موقعیت فضایی و هر کانال ویژگی در فرآیند تلفیق ارزیابی شده و ویژگی‌ها به‌صورت چندمقیاسی ادغام می‌گردند. در نهایت، تصویر تلفیقی با استفاده از رمزگشای مبتنی بر اتصال تودرتو بازسازی می‌شود. نتایج به‌دست‌آمده از آزمایش‌ها بر روی مجموعه‌داده‌های مختلف نشان می‌دهند که روش پیشنهادی، نسبت به روش‌های پیشرفته موجود، عملکرد بسیار بهتری دارد.

در ادامه پژوهش پیشین، لی و همکاران در مطالعه‌ای دیگر یکی از چالش‌های مهم روش‌های مبتنی بر یادگیری عمیق را انتخاب استراتژی مناسب برای تولید تصویر تلفیقی متناسب با کاربرد مورد نظر مطرح کرده‌اند [۲۲]. برای مواجهه با این چالش، معماری جدیدی با عنوان RFN-Nest معرفی شده است که یک شبکه end-to-end برای تلفیق تصاویر مادون‌قرمز و مرئی به‌شمار می‌آید. این معماری شامل یک شبکه تلفیق با عنوان Residual Fusion Network (RFN)است که به‌جای استفاده از روش‌های تلفیقی سنتی، از ساختار Residual بهره می‌برد. برای آموزش این مدل، دو تابع هزینه نوآورانه شامل تابع حفظ جزئیات و تابع تقویت ویژگی‌ها طراحی شده‌اند. فرایند یادگیری مدل در دو مرحله انجام می‌پذیرد: ابتدا یک Auto-Encoderبا ساختار مبتنی بر اتصال تودرتو آموزش داده می‌شود و در مرحله بعد، شبکهRFN با استفاده از توابع هزینه پیشنهادی بهینه‌سازی می‌گردد.

در یک پژوهش چندمنظوره، هان و همکاران شبکه‌ای نوین و یکپارچه با عنوان U2Fusion معرفی کرده‌اند که به‌صورت end-to-end طراحی شده و قادر است انواع مختلف تلفیق تصویر، از جمله تلفیق چندحالته (multi-modal)، چندپرده‌ای (multi-exposure) و چندفوکوس (multi-focus) را در قالب یک چارچوب واحد انجام دهد [۲۳]. این مدل با بهره‌گیری از استخراج ویژگی و اندازه‌گیری میزان اطلاعات، به‌صورت خودکار اهمیت نسبی تصاویر منبع را ارزیابی کرده و به‌صورت تطبیقی میزان حفظ اطلاعات هر تصویر را تعیین می‌کند. بدین ترتیب، نیاز به طراحی جداگانه شبکه برای هر نوع تلفیق حذف شده و فرآیند آموزش، تطبیق‌پذیر و عمومی می‌شود. همچنین، این مدل بدون نیاز به تصویر مرجع یا معیارهای از پیش تعریف‌شده، قابلیت یادگیری را حفظ می‌کند. در این پژوهش، برای ارزیابی عملکرد مدل، از مجموعه‌داده هم‌تراز تصاویر مرئی و مادون‌قرمز با عنوان RoadScene استفاده شده است. نتایج کیفی و کمی حاصل از آزمایش‌های انجام‌شده در سه نوع تلفیق، اثربخشی و جامعیت روش U2Fusion را تأیید می‌کنند.

در تمامی مطالعات بررسی‌شده، مجموعه‌داده‌هایی مورد استفاده قرار گرفته‌اند که در آن‌ها تصاویر منبع به‌صورت هم‌تراز ثبت

شده‌اند. در این مطالعه، هدف ما ایجاد یک مجموعه‌داده تلفیقی جدید با استفاده از مدل‌های مختلف تلفیق تصاویر و بر پایه داده‌های ویدئویی مجموعه‌داده NTU RGB+D است. با این حال، به‌دلیل ماهیت این مجموعه‌داده، تصاویر فروسرخ و RGB هم‌تراز نبوده و تفاوت‌های مکانی در بین فریم‌های آن‌ها وجود دارد. بنابراین، پیش از اعمال فرآیند تلفیق بر داده‌های ویدئویی، ابتدا روشی برای هم‌ترازسازی فریم‌ها پیشنهاد می‌شود تا دقت و کیفیت فرآیند تلفیق بهبود یابد. یکی از روش‌های مؤثر برای این منظور، بهره‌گیری از تکنیک‌های مبتنی بر مسائل معکوس است. در بخش بعدی، به معرفی این دسته از مسائل و ویژگی‌های آن‌ها خواهیم پرداخت.

۳- مسائل معکوس

مسائل معکوس سابقه طولانی در ریاضیات دارند و به‌طور گسترده‌ای در مهندسی برای کاربردهای مختلف استفاده می‌شوند و در سال‌های اخیر اغلب از مسائل معکوس در پردازش تصویر استفاده می‌شود [۲۵]. برای درک بهتر مفهوم مسائل معکوس، با ارائه مسائل روبه‌جلو[1] شروع می‌کنیم. به‌طور انتزاعی، اکثر سیستم‌های فیزیکی را می‌توان با مجموعه‌ای از ویژگی‌ها یا مجهولات توصیف کرد که می‌توان از آن‌ها برای استنتاج سایر ویژگی‌ها یا اندازه‌گیری‌ها استفاده نمود. به عبارت دیگر، با در نظر گرفتن معادله (۱) کمیت‌های m یک تابع ریاضی هستند که در آن f می‌تواند تصادفی یا قطعی باشد.

$$m = f(z) \tag{۱}$$

در مواردی که f خطی است، معادله (۱) را می‌توان مانند معادله (۲) توصیف کرد که به ترتیب برای موارد قطعی یا تصادفی است:

$$m = cz \qquad \text{Or } m = cz + v \tag{۲}$$

به‌طور معمول، z یک نمایش ایده‌آل از سیستم است که کامل، ساختار یافته و بدون نویز است، با این حال به‌طور متداول مقدار m ناقص، دارای نویز و بدون ساختار است بنابراین در محاسبات یک پارامتر را به عنوان مقدار خطا در نظر می‌گیریم. محاسبه مقدار m از کمیت‌های بنیادی z به عنوان یک مسئله رو به جلو شناخته می‌شود که حل این‌گونه مسائل بسیار ساده بوده زیرا مدل f برای توصیف بین کمیت‌های بنیادی و مقادیر به دست آمده موجود است.

در مقابل، وظیفه استنباط کمیت‌های بنیادی z از مقادیر مشتق شده یا اندازه‌گیری شده m به‌عنوان مسئله معکوس شناخته می‌شود. مفهوم مسائل معکوس[2] به تخمین پارامترها یا داده‌ها بر اساس مشاهدات ناکافی اشاره دارد که اغلب دارای نویز بوده و به دلیل محدودیت‌های موجود حاوی اطلاعات ناقصی در مورد پارامتر یا داده‌های هدف هستند. چنین مسائلی که در معادله (۳) نشان داده شده است به عنوان مسائل سخت شناخته می‌شوند زیرا تابع معکوس $f^{-1}()$ معمولاً ناشناخته است و فقدان اطلاعات کافی در مورد کمیت و کیفیت اندازه‌گیری‌ها بازسازی z را دشوار می‌کند.

$$z = f^{-1}(\mathrm{m}) \tag{۳}$$

برای حل چنین مسائلی ابتدا باید مشخص شود که آیا مسئله well-posed هست یا خیر. مسائل well-posed که توسط هادامارد[3] مطرح شده است به مسائلی اطلاق می‌گردد که دارای شرایط زیر باشند [۲۶]:

۱- موجودیت[4]: برای هر مشاهده m، حداقل یک مقدار مربوطه از z وجود داشته باشد.

۲- منحصربه‌فرد بودن[5]: راه حل z برای هر مشاهده m منحصربه‌فرد باشد.

۳- پیوستگی[6]: جواب z به‌طور پیوسته به m وابسته است.

اگر هر یک از شرایط ذکر شده برآورده نشود، مسئله ill-posed خواهد بود، بنابراین در راه حل پیشنهادی باید هر سه شرط هادامارد برقرار باشند. در این مطالعه راه‌حلی بر اساس مسائل معکوس برای تخمین نقاط مورد نظر برای ایجاد یک پنجره جهت برش فریم‌های RGB پیشنهاد خواهیم‌کرد. در این نوع مسائل، شرایط مختلف اغلب منجر به نویز در نتایج می‌گردد، بنابراین استفاده از راه حل‌های متناسب بسیار مهم است. روش‌های متعددی از جمله روش‌های بیزی، آماری و سایر تکنیک‌های کلاسیک برای

[1] Forward Problems
[2] Inverse Problems
[3] Hadamard
[4] Existence
[5] Uniqueness
[6] Continuity

بهبود دقت تخمین مورد استفاده قرار گرفته‌اند [۲۵]. در این مطالعه، با هدف افزایش دقت و کاهش خطا، از ویژگی‌های ماتریسی بهره خواهیم گرفت که در بخش‌های آتی به‌تفصیل به آن‌ها پرداخته خواهد شد. اما همان‌طور که پیش‌تر اشاره شد، مجموعه‌داده مورد استفاده به‌دلیل هم‌تراز نبودن داده‌ها، به‌طور مستقیم برای فرآیند تلفیق مناسب نیست. از این‌رو، پیش از معرفی روش پیشنهادی مبتنی بر مسائل معکوس برای هم‌ترازسازی، در بخش بعد به بررسی ویژگی‌ها و ساختار مجموعه‌داده NTU RGB+D خواهیم پرداخت.

٤- مجموعه داده

دیتاست NTU RGB+D بزرگ‌ترین مجموعه‌داده در زمینه تشخیص حرکت انسان بوده که در محیط کنترل شده و با استفاده از دوربین Kinect v2 جمع آوری شده است [۱۰]. اطلاعاتی که توسط این دوربین ضبط می‌شود شامل داده‌های فروسرخ، داده‌های عمق، داده‌های RGB و نیز مختصات نقاط بدنی است. داده‌های ویدئویی فروسرخ و عمق از یک سنسور گرفته می‌شوند و ابعاد آن‌ها دقیقاً مشابه است و به‌صورت ٤٢٤ × ٥١٢ است، اما داده‌های ویدئویی RGB با استفاده از سنسوری گرفته می‌شوند که در کنار سنسور فروسرخ قرار دارد و اطلاعات را با کیفیت ۱۹۲۰ × ۱۰۸۰ ذخیره می‌کند.

از سوی دیگر در فایل‌های اسکلتی مختصات و ویژگی‌های ۲۵ نقطه مفصلی اصلی بدن انسان توسط دوربین کینکت ثبت می‌گردد که محل قرارگیری آن‌ها نیز در شکل (۱) نشان داده شده است [۲٤]. این داده‌ها در فایل‌هایی با پسوند skeleton.ذخیره شده و حاوی اطلاعاتی همچون تعداد فریم‌های هر ویدئو، وضعیت track، کانفیدنس نقاط، شناسه، تعداد نقاط بدنی و ویژگی‌های هر نقطه است. هنگام ضبط اطلاعات هر نقطه، ۱۱ ویژگی مربوط به نقطه مورد نظر توسط دوربین کینکت ذخیره می‌شوند که این ۱۱ ویژگی به‌صورت زیر هستند:

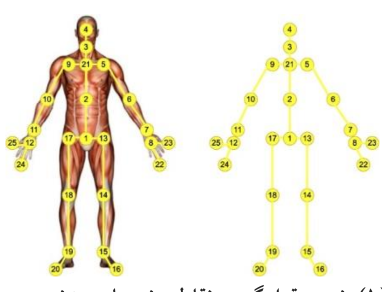
- نقاط x، y و z: این نقاط در واقع مختصات محل قرارگیری ۲۵ نقطه در فضای سه بعدی را مشخص می‌کنند.
- نقاط depth y و depth x: این نقاط بیانگر مختصات نقاط بدنی در داده‌های فروسرخ می‌باشند.
- نقاط color x و color y :این نقاط بیانگر مختصات نقاط بدنی در داده‌های RGB می‌باشند.
- نقاط X ، Y، W و Z :کینکت از کواترنیون[1] برای ارائه جهت‌گیری مفاصل استفاده می‌کند و این نقاط مربوط به محورهای خود دوربین کینکت است. در واقع، کواترنیون روشی چهار بعدی برای ذخیره‌سازی جهت‌گیری سه بعدی است.

به همراه یازده ویژگی یاد شده یک ویژگی دیگر به عنوان وضعیت track نیز همراه اطلاعات ذخیره می‌شود، که این مقدار می‌تواند سه حالت داشته باشد:

۱- اگر سوژه توسط دوربین کینکت شناسایی نشود در این حالت مقدار صفر شده و به‌طور کلی هیچ ویژگی‌ای برای هیچ یک از نقاط ذخیره نمی‌شود، این نوع داده‌ها به عنوان خطا شناخته می‌شوند و تعداد آن‌ها نسبت به کل داده‌ها بسیار ناچیز است.

۲- اگر سوژه توسط دوربین کینکت شناسایی شود ولی ردیابی آن نسبت به استاندارد دوربین ضعیف‌تر باشد و یا به عبارتی ردیابی یازده نقطه یاد شده نسبتاً ضعیف باشد، مقدار ۱ به عنوان مقدار track آن نقطه ذخیره می‌گردد.

۳- اگر سوژه توسط دوربین کینکت شناسایی شود و ردیابی تمامی یازده نقطه نیز قوی باشد مقدار ۲ به عنوان مقدار track آن نقطه ذخیره می‌گردد.



شکل (۱): نحوه قرارگیری نقاط مختصات بدنی در دوربین Kinect v2

---

[1] Quaternion

با توجه به بررسی ویژگی‌های مجموعه‌داده مورد استفاده، به‌وضوح می‌توان دریافت که این مجموعه‌داده برای تلفیق تصاویر استانداردسازی نشده است. با این حال، با در نظر گرفتن قابلیت‌های موجود، به‌ویژه در دسترس بودن مختصات نقاط بدنی در منابع داده‌ای، امکان ارائه رویکردی مبتنی بر مسائل معکوس جهت هم‌ترازسازی فریم‌های ویدئویی وجود دارد. این رویکرد در بخش بعدی به‌تفصیل مورد بررسی قرار خواهد گرفت. استانداردسازی و هم‌ترازسازی صحیح فریم‌ها، علاوه بر فراهم‌سازی امکان تلفیق دقیق داده‌ها، نقش مهمی در ارتقاء عملکرد سیستم‌های شناسایی و تشخیص حرکت انسان نیز ایفا می‌کند. در این راستا، به‌جای استخراج جداگانه ویژگی‌ها از تصاویر مادون‌قرمز یا مرئی، با ترکیب این دو منبع اطلاعاتی و بهره‌گیری از ویژگی‌های مکمل آن‌ها، می‌توان نمایش‌های ویژگی دقیق‌تر و غنی‌تری تولید کرد که منجر به بهبود قابل‌توجه دقت در سامانه‌های تشخیص حرکت انسان خواهد شد [۲۷،۲۸،۲۹].

۵- روش پیشنهادی

همان‌طور که اشاره شد، در این پژوهش سعی داریم تا با استفاده از تکنیک‌های تلفیق تصاویر داده‌های فروسرخ و RGB که توسط دوربین Kinect v2 ضبط می‌شوند را با یکدیگر تلفیق نموده و در نتیجه داده‌های بهبود یافته و جدیدی تولید کرد که حاوی اطلاعات فشرده و بالایی از هر دو نوع داده ویدئویی ذکر شده هستند. اولین شرط برای تلفیق ویدئوهای فروسرخ و RGB، تراز بودن آن‌ها است که در قسمت‌های قبل به آن اشاره گردید، پس از آماده‌سازی داده‌ها باید یکی از تکنیک‌های تلفیق جهت ترکیب ویدئوها انتخاب گردد.

در سال‌های اخیر با رشد کاربردهای تلفیق اطلاعات روش‌های زیادی در این زمینه به خصوص تلفیق تصاویر چند سنسوری معرفی شده است، اما با رشد یادگیری عمیق و با توجه به نتایج تحقیقات صورت گرفته می‌توان به تأثیر غیر قابل انکار آن در افزایش دقت تلفیق تصاویر اشاره نمود لذا در این پژوهش به دنبال پیاده‌سازی و استفاده از مدل‌هایی هستیم که از روش‌های یادگیری عمیق برای تلفیق تصاویر استفاده می‌کنند [۳۰].

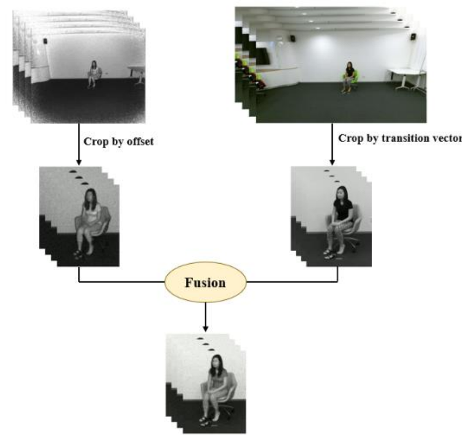
همان‌طور که اشاره شد مجموعه داده NTU RGB+D مخصوص تشخیص حرکت انسان بوده لذا برای تلفیق تصاویر استانداردسازی نشده است و ابعاد دو داده ویدئویی فروسرخ و RGB با یکدیگر متفاوت است. با توجه به مشکل یاد شده می‌بایست دو نوع داده یکسان شوند، به‌گونه‌ای که هر دو داده کاملاً تراز بوده و فضای همسانی را پوشش دهند.

(ب) فروسرخ **(الف) RGB**

شکل (۲): فریم‌های یکسان از دو داده ویدئویی فروسرخ و RGB

فریم‌های مشابهی در شکل (۲) از یک نمونه داده ویدئویی از دیتاست نشان داده شده است، همان‌طور که مشاهده می‌کنید باوجود اینکه دو تصویر از سوژه یکسانی بوده و تنظیمات دوربین کاملاً یکسانی دارند، اما در فضایی که پوشش می‌دهند با یکدیگر متفاوت هستند. برای حل این مشکل روشی ارائه خواهیم‌داد که با استفاده از آن تصاویر فروسرخ و RGB را به‌گونه‌ای برش می‌دهیم که هر دو داده ویدئویی کاملاً تراز شوند و محیط یکسانی از صحنه را پوشش دهند. در نهایت، روش پیشنهادی بر روی تمامی نمونه‌های مجموعه داده NTU RGB+D اجرا می‌شود تا مجموعه‌ای کامل از داده‌های فروسرخ و RGB به دست آید.

در این مجموعه، هر نمونه فروسرخ به‌طور دقیق با نمونه متناظر RGB خود تراز شده است، به گونه‌ای که امکان مقایسه، تحلیل و تلفیق دقیق اطلاعات میان این دو نوع داده با کیفیت و دقت بالا فراهم شود. شکل (۳) شمای کلی روش پیشنهادی را نشان می‌دهد. در ادامه رویکردی که برای برش دو نوع داده ویدئویی طراحی شده را ارائه خواهیم‌کرد.

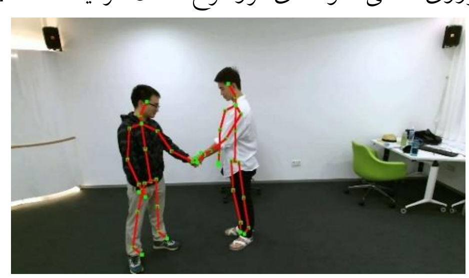


شکل (۳): شمای کلی روش پیشنهادی

۵-۱ برش ویدئوهای فروسرخ

با توجه به وجود مختصات قرارگیری نقاط بدنی در داده‌های فروسرخ می‌توان از آن‌ها برای برش[1] ویدئوها استفاده نمود. به این منظور ابتدا باید مقادیر $xmin$ ،$xmax$ ،$ymin$ و $ymax$ در میان نقاط مختصات بدنی موجود در تمام فریم‌های یک نمونه ویدیویی یافت شوند.

در شکل (٤) محل قرارگیری نقاط مختصات بدنی بر روی بدن سوژه در یک فریم مشخص است، همان‌طور که مشاهده می‌کنید برخی نقاط مانند سر، دست‌ها و پاها نقاطی هستند که احتمال انتخاب شدن آن‌ها به عنوان کمترین و یا بیشترین مقدار بسیار بالا است اما اگر برش از این نقاط صورت گیرد آنگاه ممکن است تمام بدن سوژه در پنجره کراپ شده قرار نگیرد، بنابراین برای حل این مشکل ما مقداری را به عنوان آفست که برابر ۲۰ پیکسل است، در محاسبات به نقاط ماکسیمم و مینیمم اضافه می‌کنیم. این روش برروی تمامی نمونه‌های فروسرخ اعمال گردیده است [۳۱].



شکل (٤): وضعیت قرارگیری نقاط اسکلتی بر بدن انسان در نمونه داده‌ای RGB [۱۰]

[1] Cropping

۵-۲ برش ویدئوهای RGB

با توجه به اطلاعات موجود در فایل‌های skeleton و مشخص بودن مختصات (x, y) یک نقطه در داده فروسرخ و RGB می‌توان با استفاده از روش‌های ریاضی و مسائل معکوس[۲۵]تبدیلی پیاده‌سازی کرد که با استفاده از آن می‌توانیم مختصات یک نقطه در داده فروسرخ را به مختصات همان نقطه در RGB تبدیل کرده و براساس آن نقاط جدید پنجره‌ای تنظیم و کراپ کنیم.

برای درج مختصات دقیق نقاط، باید آن‌ها را در اندازه پیکسل نقاط نیز ضرب کنیم [۳۲]. در دوربین Kinect v2 اندازه پیکسل‌ها در داده‌های RGB برابر با$h_{RGB} = w_{RGB} = 6.0\ (mm)$ و در داده‌های فروسرخ $3.38\ (mm)$ و $w_{IR} = 5.12\ (mm)$ $h_{IR} = 4.24\ (mm)$ است. در ادامه به توضیح روش طراحی شده خواهیم پرداخت. برای یک تصویر در عمق ثابت ماتریسی به‌صورت زیر داریم [۳۳]:

$$x = K[R\ T]\ X \tag{۱}$$

با توجه به ماتریس بالا K نشان دهنده پارامترهای درونی دوربین مورد نظر و پارامترهای R و T مقادیر چرخش و برگردان (پارامترهای بیرونی) هستند، لذا خواهیم داشت:

$$w\begin{bmatrix}u\\v\end{bmatrix} = \begin{bmatrix}\alpha & s & u_0\\0 & \beta & v_0\\0 & 0 & 1\end{bmatrix}\begin{bmatrix}r_{11} & r_{12} & r_{13} & t_x\\r_{21} & r_{22} & r_{23} & t_y\\r_{31} & r_{32} & r_{33} & t_z\end{bmatrix}\begin{bmatrix}x\\y\\z\end{bmatrix} \tag{۲}$$

در رابطه (۲)، α و β مقدار فاصله کانونی، s نشان‌دهنده مقدار skew، u_0 و v_0 مقادیر principal points، r و t مقادیر چرخش و برگردان هستند، همچنین x، y و z مختصات نقطه در دنیای واقعی و u و v مختصات نقطه در تصویر می‌باشند. حال فرض می‌کنیم نقطه‌ای در تصویر فروسرخ به مختصات $[u_{IR}\quad v_{IR}]$و در تصویر RGB به مختصات $[u_{RGB}\quad v_{RGB}]$ قرار گرفته‌باشد، لذا برای هر یک از تصاویر IR و RGB خواهیم داشت:

$$\begin{bmatrix}u_{IR}\\v_{IR}\end{bmatrix} = k_{IR}[R_{IR}\quad t_{IR}]\begin{bmatrix}x\\y\\z\end{bmatrix} \tag{۳}$$

$$\begin{bmatrix}u_{RGB}\\v_{RGB}\end{bmatrix} = k_{RGB}[R_{RGB}\quad t_{RGB}]\begin{bmatrix}x\\y\\z\end{bmatrix}$$

سپس با استفاده از عملیات جابه‌جایی، در معادله (٤) داریم:

$$\begin{bmatrix}u_{IR}\\v_{IR}\end{bmatrix} = (k_{IR}[R_{IR}\quad t_{IR}])\,(k_{RGB}[R_{RGB}\quad t_{RGB}])^{-1} \Rightarrow \begin{bmatrix}u_{IR}\\v_{IR}\end{bmatrix} = A_{2*2}\begin{bmatrix}u_{RGB}\\v_{RGB}\end{bmatrix} \quad (٤)$$

با توجه به آنکه مقادیر پارامترهای داخلی و خارجی دوربین در دسترس نیست لذا فرض می‌کنیم ماتریس A نشان‌دهنده تمام پارامترهای درونی و بیرونی است، لذا می‌توان با تخمین این ماتریس، تبدیل بین دو تصویر فروسرخ و RGB را به دست آورده و با استفاده از آن برش دقیق‌تری از دو نوع داده انجام داد؛ و در ادامه خواهیم داشت:

$$\begin{bmatrix}u_{IR}\\v_{IR}\end{bmatrix} = A_{2*2}\begin{bmatrix}u_{RGB}\\v_{RGB}\end{bmatrix} \Rightarrow \begin{bmatrix}u_{IR}\\v_{IR}\end{bmatrix} = \begin{bmatrix}a_{11} & a_{12}\\a_{21} & a_{22}\end{bmatrix}\begin{bmatrix}u_{RGB}\\v_{RGB}\end{bmatrix} \quad (٥)$$

لذا با استفاده از تساوی بالا می‌توان معادلات زیر را به دست آورد:

$$\begin{cases}u_{IR} = a_{11}\,u_{RGB} + a_{12}\,v_{RGB}\\v_{IR} = a_{21}\,u_{RGB} + a_{22}\,v_{RGB}\end{cases} \quad (٦)$$

با توجه به آنکه تمامی مقادیر ماتریس A مجهول هستند پس برای به دست آوردن این ٤ مقدار حداقل به دو نقطه احتیاج خواهیم‌داشت، بنابراین با استفاده از دو نقطه موجود از داده‌های مختصات بدنی و ساخت ٤ معادله می‌توان مقادیر $(a_{11}, a_{22}, a_{21}, a_{22})$ از ماتریس A را به دست آورد.

$$\begin{bmatrix}u_{IR1} & u_{IR2}\\v_{IR1} & v_{IR2}\end{bmatrix} = \begin{bmatrix}a_{11} & a_{12}\\a_{21} & a_{22}\end{bmatrix}\begin{bmatrix}u_{RGB1} & u_{RGB2}\\v_{RGB1} & v_{RGB2}\end{bmatrix} \quad (٧)$$

و لذا برای دو نقطه مختصات بدنی انسان در فریم‌های فروسرخ و RGB معادلات زیر را داریم:

$$\begin{cases}u_{IR1} = a_{11}\,u_{RGB1} + a_{12}\,v_{RGB1}\\v_{IR1} = a_{21}\,u_{RGB1} + a_{22}\,v_{RGB1}\end{cases} \quad \begin{cases}u_{IR2} = a_{11}\,u_{RGB2} + a_{12}\,v_{RGB2}\\v_{IR2} = a_{21}\,u_{RGB2} + a_{22}\,v_{RGB2}\end{cases} \quad (٨)$$

پس با توجه به ضرب ماتریس‌ها خواهیم داشت:

$$Y = CZ$$

$$\begin{bmatrix}u_{IR1}\\v_{IR1}\\u_{IR2}\\v_{IR2}\end{bmatrix} = \begin{bmatrix}u_{RGB1} & v_{RGB1} & 0 & 0\\0 & 0 & u_{RGB1} & v_{RGB1}\\u_{RGB2} & v_{RGB2} & 0 & 0\\0 & 0 & u_{RGB2} & v_{RGB2}\end{bmatrix}\begin{bmatrix}a_{11}\\a_{12}\\a_{21}\\a_{22}\end{bmatrix} \quad (٩)$$

با توجه به اینکه مقادیر دو ماتریس Y و C در دسترس هستند، با جایگذاری این مقادیر، بردار Z به دست خواهد آمد. این بردار نشان‌دهنده تبدیل هر نقطه از تصویر فروسرخ به نقطه متناظر آن در تصویر RGB است. به عبارت ساده‌تر، با استفاده از بردار Z می‌توان موقعیت هر نقطه در تصویر فروسرخ را به مکان دقیق آن در تصویر RGB نگاشت کرد. بنابراین، با استفاده از مختصات جدیدی که به‌دست آمده‌اند، می‌توان عملیات برش متناظر را روی فریم‌های RGB متناظر با فریم‌های فروسرخ آن انجام داد. این فرآیند گام مهمی در آماده‌سازی داده‌ها برای انجام تلفیق مؤثر و دقیق بین دو نوع تصویر محسوب می‌شود. در شکل (٥) یک نمونه فریم کراپ شده را مشاهده می‌کنید.

RGB (الف) — (ب) فروسرخ

شکل (٥): دو فریم کراپ شده با استفاده از روش پیشنهادی

نکته قابل توجه در این روش این است که می‌توان با افزایش تعداد نقاط و با استفاده از خاصیت ماتریس‌ها میزان خطا در بردار Z را کاهش داد. این روش برروی تمامی داده‌های RGB پیاده‌سازی گردیده است. مراحل انجام روش پیشنهادی که پیش‌تر توضیح دادیم، به‌صورت گام‌به‌گام در الگوریتم (١) ارائه شده است.

الگوریتم (۱) : روش پیشنهادی جهت برش فریم‌های ویدئویی داده‌های RGB

ورودی: ویدئوهای RGB ، ویدئوهای فروسرخ، داده‌های اسکلتون، نام نمونه‌ها

۱- برای هر نمونه در مجموعه داده مراحل زیر انجام می‌گردد:
- وضعیت ردیابی نقاط مفصلی در هر نمونه بررسی می‌شود.
- اگر وضعیت ردیابی به‌طور دقیق انجام شده باشد:
  - آن نمونه به لیست $high\_track\_state$ اضافه می‌شود.

۲- مقدار اولیه z_total برابر با صفر در نظر گرفته می‌شود.

۳- برای هر نمونه در لیست $high\_track\_state$:
- ماتریس c_{ij} با مقادیر مختصات نقاط مفصلی از ویدئوهای فروسرخ تنظیم می‌شود.
- ماتریس y_{jk} با مقادیر مختصات نقاط مفصلی از ویدئوهای RGB تنظیم می‌شود.
- ماتریس z به‌صورت $z = (c^T * c)^{-1} * (c^T * y)$ محاسبه می‌شود.
- z_total به‌صورت $(z\_total + z) / (lenght[high\_track\_state])$ محاسبه می‌شود.

۴- به دست آوردن ماتریس انتقال برای هر نمونه در مجموعه:
- ماتریس p_{mn} با مقادیر $(x_{min}\_ir(seq), y_{min}\_ir(seq))$ و $(x_{max}\_ir(seq), y_{max}\_ir(seq))$ تنظیم می‌گردد.
- ماتریس m به‌صورت محاسبه $\mathrm{m} = (z\_total * p^T)$ می‌شود.
- فریم‌های ویدئوی RGB با استفاده از ماتریس انتقال m برش می‌خورند.

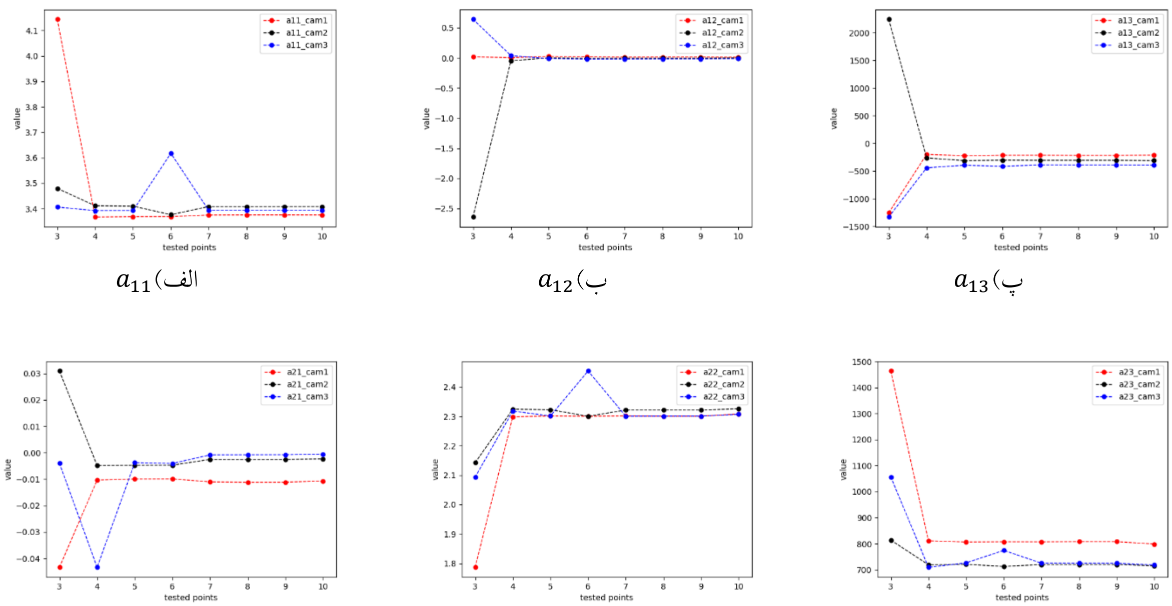
خروجی: داده‌های ویدئویی RGB برش‌خورده.

۳-۵ کاهش خطا

همان‌طور که در بخش (٤) ذکر شد، یکی از شروط اصلی برای well-posed بودن یک مسئله شرط موجودیت است. اما از آنجایی که هیچ پاسخ صریحی برای مسئله وجود ندارد و داده‌ها دارای نویز هستند، شرایط قبلی برآورده نمی‌شود و در نتیجه شرط موجودیت دچار شکست شده و نقض می‌شود [۲۵]. از آنجایی که پاسخ دقیقی وجود ندارد، ما به دنبال بهترین تخمین هستیم، با این حال برآورد پاسخ همیشه دارای خطا خواهد بود، بنابراین در این روش ما به دنبال بهترین پاسخ با کمترین میزان خطا هستیم.

برای حل چنین مسائلی از راه حل‌های مختلفی از جمله روش‌های بیزی، عملیات آماری و... استفاده می‌شود. در این تحقیق کوشش بر آن است که با استفاده از روش‌های آماری و با ثابت نگه داشتن تعداد معادلات و افزایش تعداد نقاط، خطا را کاهش دهیم. معادلات ارائه شده در بخش قبل را می‌توان با استفاده از حداقل سه نقطه مختصات بدنی حل کرد، بنابراین پیشنهاد ما برای کاهش خطا افزایش تعداد نقاط درگیر در معادلات و ثابت نگه داشتن تعداد معادلات است.

شکل (٦): مقادیر مولفه‌های بردار z حاصله با استفاده از تعداد نقاط مختلف (از ۳ تا ۱۰ نقطه) درگیر در معادلات بر روی داده‌های ویدئویی دوربین ۱، دوربین ۲ و دوربین ۳



الف) a_{11} ب) a_{12} پ) a_{13}

ت) a_{21} ث) a_{22} ج) a_{23}

لذا با افزایش تعداد نقاط و ثابت نگه داشتن تعداد معادلات می‌توان خطای بردار Z را کاهش داده و سپس با استفاده از خواص ماتریس‌ها مقدار بردار Z را محاسبه کرد. با توجه به رویکرد اشاره شده، به منظور کاهش خطا در روش پیشنهادی تعداد نقاط را افزایش داده و در نتیجه با افزایش تعداد نقاط، شاهد کاهش نوسانات مقادیر بردار حاصله بوده‌ایم. با بررسی نتایج نشان داده شده در شکل (٦) می‌توان مشاهده نمود که افزایش تعداد نقاط از سه به شش نقطه منجر به یک پاسخ تخمینی بهتر و کاهش نوسانات ناشی از نویز می‌شود، اما با افزایش تعداد نقاط به بیشتر از شش نقطه، هیچ پیشرفت دیگری قابل مشاهده نیست. بدین ترتیب با توجه به تست‌های انجام شده بر روی فریم‌های ویدئویی، تصمیم گرفته شد از ٦ نقطه با معادلات جاری استفاده شود.

٦- ارزیابی و نتایج

در این بخش، عملکرد روش پیشنهادی ارائه می‌شود. تمامی آزمایش‌ها روی سیستمی مجهز به پردازنده Intel Core i7-10750H و کارت گرافیک NVIDIA GeForce GTX 1650 Ti (4GB)انجام شده است. پیاده‌سازی با استفاده از زبان برنامه‌نویسی Python وکتابخانه‌های مرتبط صورت گرفته و اجرای الگوریتم بر روی کل مجموعه داده، با بهره‌گیری از GPU، حدود ۱۵۰ ساعت به طول انجامیده است. پس از معرفی مجموعه داده HDF5 و فرایند پیش‌پردازش داده‌ها، روش پیشنهادی بر اساس نتایج حاصل از رویکردهای تلفیقی مختلف ارزیابی خواهد شد.

٦-١ پیش‌پردازش داده‌ها

پردازش مجموعه داده NTU RGB+D به دلیل حجم بالا و فرمت ویدئویی داده‌های فروسرخ و RGB چالش‌برانگیز است. لذا برای بهبود مدیریت و افزایش سرعت پردازش، داده‌ها را به فرمت‌های مناسب ذخیره‌سازی و پردازش ویدئو تبدیل می‌کنیم. فرمت HDF5یکی از گزینه‌های مناسب در این زمینه است که داده‌ها را به‌صورت سلسله‌مراتبی سازمان‌دهی می‌کند. این ویژگی امکان جست‌وجوی سریع و سازگاری با انواع داده‌ای را فراهم می‌آورد [٣٤].

یکی از مزایای کلیدی HDF5 این است که می‌توان تنها داده مورد نیاز را بدون بارگذاری کل مجموعه فراخوانی کرد. اگرچه تبدیل ویدئوها به این فرمت زمان‌بر است، اما روشی کارآمد برای مدل ما محسوب می‌شود. بنابراین، پیش از اجرای مدل، ویدئوهای فروسرخ، RGB و داده‌های مختصات بدنی را به فرمت H5 تبدیل کرده‌ایم.

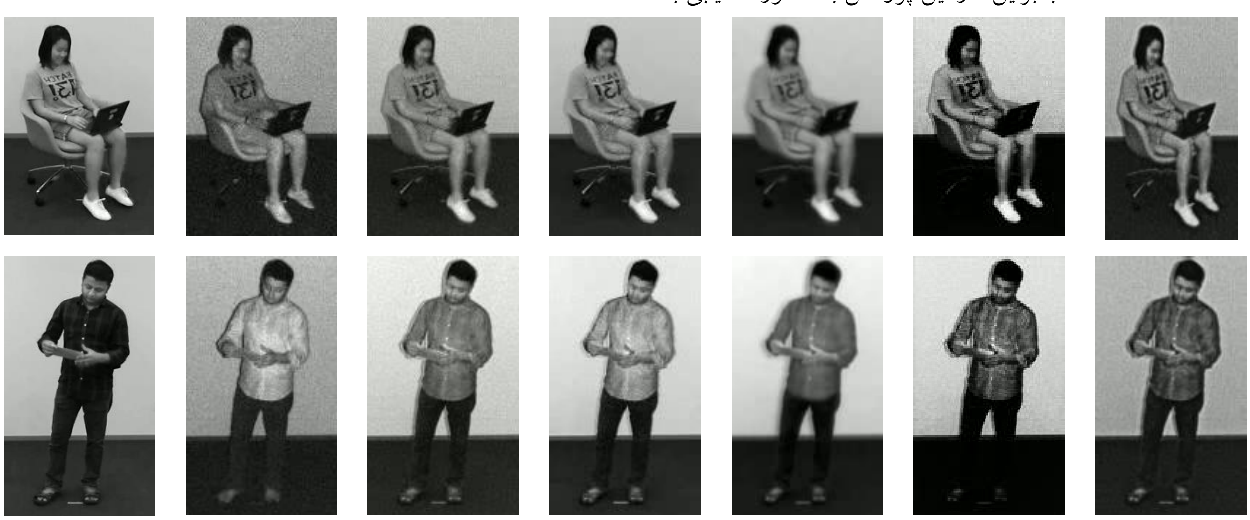
٦-٢ ارزیابی نتایج

برای بررسی کیفیت روش پیشنهادی، ابتدا داده‌های برش داده شده فروسرخ و RGB با استفاده از مدل‌های مختلف تلفیق تصاویر از جمله Dense Fuse [١٩]، Nest Fuse [٢١]، RFN-Nest [٢٢]، DID Fuse [٢٠] و U2 Fusion [٢٣] با یکدیگر تلفیق شده‌اند. این مدل‌ها هر یک به نحوی خاص، ویژگی‌ها و مزایای خود را در فرآیند تلفیق تصاویر به کار می‌برند و سعی دارند تا بهترین نتایج را از لحاظ حفظ جزئیات و بهبود کیفیت خروجی ارائه دهند.

مدل‌های انتخاب‌شده برای تولید داده‌های تلفیقی، مدل‌هایی هستند که ساختار آن‌ها بر پایه روش‌های یادگیری عمیق طراحی شده و در حوزه تلفیق اطلاعات در سطح پیکسل، عملکرد بسیار مطلوب و قابل‌توجهی به خصوص در تلفیق داده‌های فروسرخ و تصاویر RGB نشان داده‌اند. بنابراین، در این پژوهش به منظور دستیابی به نتایجی با کیفیت بالا و تولید داده‌های تلفیقی دارای ویژگی‌های قوی و قابل‌اعتماد، این مدل‌ها به‌صورت ویژه انتخاب و به کار گرفته‌شده‌اند تا ضمن افزایش دقت، کیفیت نهایی داده‌های تلفیقی بهبود یابد.

پس از تلفیق، نتایج به‌دست‌آمده با استفاده از متریک‌های مختلف همچون EN ، MI، SSIM و MS-SSIM مورد بررسی و ارزیابی قرار گرفته‌اند [٣٥]. این متریک‌ها برای سنجش کیفیت تصویر و ارزیابی دقت روش‌های تلفیق، ابزارهای مناسبی هستند که به‌طور دقیق و علمی قابلیت‌های هر روش را سنجیده و نتایج را به صورت کمی ارزیابی می‌کنند.

نمونه‌هایی از فریم‌های تلفیق‌شده از نمونه ویدئوهای مختلف مجموعه‌داده در شکل‌های (٧) و (٨) ارائه شده‌اند تا مقایسه بصری بین مدل‌های مختلف انجام شود. همان‌طور که در این شکل‌ها مشاهده می‌شود، این خروجی‌ها براساس مدل‌های ذکر شده که هرکدام مشخصه‌ها و ویژگی‌های متفاوتی دارند، به‌دست‌آمده‌اند. نتایج نشان می‌دهند که کیفیت خروجی از لحاظ بصری بسیار مطلوب بوده است. این نتایج ازنظر وضوح تصویر و بهبود جزئیات، نشان‌دهنده عملکرد مناسب روش پیشنهادی در تراز کردن دو نوع داده‌ای است که در نتیجه آن، تصاویر تلفیق‌شده با کیفیت بالا و جزئیات دقیق ایجاد شده‌اند.



RGB cropped | **IR cropped** | **Dense Fuse** | **Nest Fuse** | **RFN-Nest Fuse** | **DID Fuse** | **U2 Fusion**

شکل (٧): نمونه‌هایی از فریم‌های تلفیق شده از ویدئوهای تک نفره

RGB cropped | IR cropped | Dense Fuse | Nest Fuse | RFN-Nest Fuse | DID Fuse | U2 Fusion

شکل (۸): نمونه‌هایی از فریم‌های تلفیق شده از ویدئوهای دو نفره

برای ارزیابی بیشتر، از معیارهای EN و MI برای سنجش میزان اطلاعات موجود و همپوشانی تصاویر منبع در تصویر خروجی، و از معیارهای SSIM و MS-SSIM برای اندازه‌گیری شباهت و کیفیت ساختاری تصاویر تولیدی استفاده شده است. نتایج حاصل از مقیاس‌های کمی بحث شده در جدول (۱) نشان داده شده است. با توجه به بالاترین مقادیر حاصله از معیارهای EN و MI که به ترتیب ۷/۱ و ۱۳/۰۱ می‌باشند، این مقادیر نشان‌دهنده حجم بالای اطلاعات انتقالی به تصاویر تلفیقی ساخته‌شده است. به عبارت دیگر، این مقادیر بالا بیانگر آن هستند که فرآیند تلفیق، اطلاعات مهم را از هر دو نوع داده (فروسرخ و RGB) به‌طور مؤثر حفظ کرده و به تصاویر تلفیقی منتقل کرده است.

از سوی دیگر، با توجه به بالاترین مقادیر به‌دست‌آمده از دو متریک SSIMو MS-SSIM که به ترتیب برابر با ۰/۷۸ و ۰/۸٤ هستند، می‌توان نتیجه گرفت که کیفیت ساختاری تصویر تولیدی بسیار بالا است. این مقادیر نشان‌دهنده حفظ ویژگی‌های ساختاری مهم و شباهت بالای تصویر تولیدی به تصویر اصلی است. بنابراین، با توجه به نتایج به‌دست‌آمده از معیارهای مختلف، می‌توان ادعا کرد که روش طراحی‌شده به‌طور مؤثر عمل کرده است. این روش توانسته است برش‌های باکیفیتی از داده‌های فروسرخ و RGB ارائه دهد، به‌طوری‌که تصاویر منبع به‌طور دقیق تراز شده و به ورودی مدل تلفیق داده شده‌اند و خروجی حاصله نیز با توجه به مقادیر متریک‌ها از کیفیت بسیار خوبی برخوردار است.

جدول (۱): نتایج ارزیابی ویدئوهای تلفیق شده با استفاده از متریک‌های منتخب

	EN	MI	SSIM	MS-SSIM
Dense Fuse	۷/۱۷٦٤٥	۱۲/٦٥٨٦٩	۰/۷۸۱٤٥	۰/۸٤۷۲٥
Nest Fuse	۷/۱۱۲۳٤	۱۳/۰۱۰٤٤	۰/۷۸٥۸٥	۰/۸۳۲۸٦
RFN-Nest Fuse	۷/۱۱٦٥۰	۱۲/۸۱٤۹٥	۰/۷٥۰٦۸	۰/۸۳۱۸۱
DID Fuse	٦/۲۲۹٦۸	۱۱/۸٤۸۳۷	۰/٥۰٦٦٥	۰/۸۰٦۱٥
U2 Fusion	٦/٦۳٦٦۰	۱۱/۸۸۱٤۲	۰/۷۷٥٦۹	۰/۸٤۳۲٦

با توجه به نتایج به دست آمده حاصل از متریک‌های EN، MI، SSIM و MS-SSIM در جدول (۱) و همچنین نمونه فریم‌های تلفیق شده با استفاده از مدل‌های اشاره شده در شکل (۷) و شکل (۸) می‌توان ادعا کرد که روش پیشنهادی جهت تراز کردن داده‌های مجموعه داده NTU RGB+D نتایج بسیار خوبی را به همراه داشته است. لذا می‌توان از روش پیشنهادی برای تراز کردن داده‌های ویدئویی فروسرخ و RGB استفاده نموده و مجموعه داده تلفیقی جدیدی به وجود آورد. با توجه به کیفیت حاصله از این مجموعه داده می‌تواند در اهداف مطالعاتی همچون مدل‌های تلفیق تصاویر و سامانه‌های تشخیص حرکت انسان به ما کمک کند.

۷– نتیجه‌گیری

در این مطالعه روشی برای تراز و برش داده‌های ویدئویی فروسرخ و RGB از مجموعه داده NTU RGB+D ارائه گردید. با استفاده از روش پیشنهادی، داده‌های فروسرخ و RGB با یکدیگر تراز شده سپس با استفاده از تکنیک‌های تلفیق مختلف مجموعه داده تلفیق شده جدیدی به منظور دستیابی به هدف خود یعنی ارائه داده‌های بهبود یافته برای مدل تشخیص حرکت انسان ارائه شد. با توجه به نتایج به دست آمده حاصل از معیارهای مختلف کمی بر روی روش پیشنهادی می‌توان دریافت که تراز و برش داده‌های ویدئویی عملکرد خوبی داشته است. بنابراین می‌توان با استفاده از این روش تمامی نمونه‌های موجود در مجموعه داده را پردازش نموده تا بتوان مجموعه داده تلفیق شده جدیدی ساخت و سپس از آن برای اهداف آینده یعنی استفاده به عنوان ورودی سامانه‌های تشخیص حرکت انسان استفاده کرد.

یکی از چالش‌های این روش، درصد خطایی است که به دلیل ثبت داده‌ها توسط دو سنسور متفاوت دوربین Kinect V2 به‌وجود می‌آید. این خطا اجتناب‌ناپذیر است، اما در آینده می‌توان با استفاده از روش‌هایی مانند یادگیری عمیق، دقت را بهبود بخشید. از سوی دیگر، با توجه به حجم بالای مجموعه‌داده‌ها، پردازش و استفاده از این داده‌ها در مدل‌های پیچیده با چالش‌های متعددی همراه است. این چالش‌ها شامل نیاز به منابع محاسباتی بالا، زمان پردازش زیاد و پیچیدگی‌های الگوریتمی است. با در نظر گرفتن این محدودیت‌ها، می‌توان ساختارهایی بهینه‌تر طراحی و به کار گرفت. اما با توجه به تعداد بالای داده‌ها، می‌توان به نتایج به‌دست‌آمده اعتماد بیشتری داشت؛ چرا که داده‌های بیشتر موجب تقویت قابلیت تعمیم مدل‌ها به شرایط مختلف و افزایش دقت پیش‌بینی‌ها می‌شود. بنابراین، مجموعه‌داده تلفیقی تولیدشده در این مطالعه قابلیت استفاده در پژوهش‌های آتی در زمینه‌هایی نظیر مدل‌های تلفیقی، تشخیص حرکت، شناسایی اشیاء، تحلیل و پردازش ویدئو و سایر کاربردهای بینایی ماشین را داراست.

References

[1] K. Rani and R. Sharma, “Study of different image fusion algorithm,” Int. J. Emerg. Technol. Adv. Eng., vol. 3, no. 5, pp. 288–291, May 2013.

[2] D. Mishra and B. Palkar, “Image fusion techniques: a review,” Int. J. Comput. Appl., vol. 130, no. 9, pp. 7–13, 2015, doi: 10.5120/ijca2015907084.

[3] J. Ma, Y. Ma, and C. Li, “Infrared and visible image fusion methods and applications: A survey,” Inf. Fusion, vol. 45, pp. 153–178, 2019, doi: 10.1016/j.inffus.2018.02.004.

[4] S. Li, X. Kang, L. Fang, J. Hu, and H. Yin, “Pixel-level image fusion: A survey of the state of the art,” Inf. Fusion, vol. 33, pp. 100–112, 2017, doi: 10.1016/j.inffus.2016.05.004.

[5] D. E. Nirmala and V. Vaidehi, “Comparison of Pixel-level and feature level image fusion methods,” in Proc. 2nd Int. Conf. Comput. Sustain. Global Dev. (INDIACom), New Delhi, India, Mar. 2015, pp. 743–748.

[6] G. Xiao, D. P. Bavirisetti, G. Liu, and X. Zhang, “Decision-level image fusion,” Image Fusion, pp. 149–170, 2020.

[7] R. Poppe, “A survey on vision-based human action recognition,” Image Vis. Comput., vol. 28, no. 6, pp. 976–990, 2010, doi: 10.1016/j.imavis.2009.11.014.

[8] M. Karim, S. Khalid, A. Aleryani, J. Khan, I. Ullah, and Z. Ali, "Human action recognition systems: A review of the trends and state-of-the-art," IEEE Access, 2024, doi: 10.1109/ACCESS.2024.3373199.

[9] X. Jin, Q. Jiang, S. Yao, D. Zhou, R. Nie, J. Hai, and K. He, "A survey of infrared and visual image fusion methods," Infrared Phys. Technol., vol. 85, pp. 478-501, 2017, doi: 10.1016/j.infrared.2017.07.010.

[10] Shahroudy, J. Liu, T. T. Ng, and G. Wang, "Ntu rgb+d: A large scale dataset for 3d human activity analysis," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 1010-1019, doi: 10.1109/CVPR.2016.115.

[11] He, Q. Liu, H. Li, and H. Wang, "Multimodal medical image fusion based on IHS and PCA," Procedia Eng., vol. 7, pp. 280-285, 2010, doi: 10.1016/j.proeng.2010.11.045.

[12] Lu, C. Miao, and H. Wang, "Pixel level image fusion based on linear structure tensor," in 2010 IEEE Youth Conf. Inf., Comput. Telecommun., 2010, pp. 303-306.

[13] U. Patil and U. Mudengudi, "Image fusion using hierarchical PCA," in 2011 Int. Conf. Image Inf.

Process., 2011, pp. 1-6, doi: 10.1109/ICIIP.2011.6108966.

[14] W. He, W. Feng, Y. Peng, Q. Chen, G. Gu, and Z. Miao, "Multi-level image fusion and enhancement for target detection," Optik, vol. 126, no. 11–12, pp. 1203-1208, 2015, doi: 10.1016/j.ijleo.2015.02.092.

[15] Y. Liu, X. Chen, H. Peng, and Z. Wang, "Multi-focus image fusion with a deep convolutional neural network," Inf. Fusion, vol. 36, pp. 191-207, 2017, doi: 10.1016/j.inffus.2016.12.001.

[16] Z. Ahmad, A. Tabassum, L. Guan, and N. Khan, "ECG heart-beat classification using multimodal image fusion," in ICASSP 2021-2021 IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), 2021, pp. 1330-1334, doi: 10.1109/ICASSP39728.2021.9414709.

[17] L. Tang, X. Xiang, H. Zhang, M. Gong, and J. Ma, "DIVFusion: Darkness-free infrared and visible image fusion," Inf. Fusion, vol. 91, pp. 477-493, 2023, doi: 10.1016/j.inffus.2022.10.034.

[18] Y. Chen, L. Cheng, H. Wu, F. Mo, and Z. Chen, "Infrared and visible image fusion based on iterative differential thermal information filter," Opt. Lasers Eng., vol. 148, p. 106776, 2022, doi: 10.1016/j.optlaseng.2021.106776.

[19] H. Li and X. J. Wu, "DenseFuse: A fusion approach to infrared and visible images," IEEE Trans. Image Process., vol. 28, no. 5, pp. 2614-2623, 2019, doi: 10.1109/TIP.2019.2899946.

[20] Z. Zhao, S. Xu, C. Zhang, J. Liu, P. Li, and J. Zhang, "DIDFuse: Deep image decomposition for infrared and visible image fusion," arXiv preprint arXiv:2003.09210, 2020, doi: 10.24963/ijcai.2020/135.

[21] H. Li, X. J. Wu, and T. Durrani, "NestFuse: An infrared and visible image fusion architecture based on nest connection and spatial/channel attention models," IEEE Trans. Instrum. Meas., vol. 69, no. 12, pp. 9645-9656, 2020, doi: 10.1109/TIM.2020.3005230.

[22] H. Li, X. J. Wu, and J. Kittler, "RFN-Nest: An end-to-end residual fusion network for infrared and visible images," Inf. Fusion, vol. 73, pp. 72-86, 2021, doi: 10.1016/j.inffus.2021.02.023.

[23] H. Xu, J. Ma, J. Jiang, X. Guo, and H. Ling, "U2Fusion: A unified unsupervised image fusion network," IEEE Trans. Pattern Anal. Mach. Intell., vol. 44, no. 1, pp. 502-518, 2020, doi: 10.1109/TPAMI.2020.3012548.

[24] R. Dang, C. Liu, M. Liu, and Q. Chen, “Channel attention and multi-scale graph neural networks for skeleton-based action recognition,” AI Commun., vol. 35, no. 3, pp. 187–205, 2022, doi: 10.3233/AIC-210250.

[25] P. Fieguth, Statistical Image Processing and Multidimensional Modeling. Springer, 2010, doi: 10.1007/978-1-4419-7294-1.

[26] J. Hadamard, Lectures on Cauchy’s Problem in Linear Partial Differential Equations, vol. 15. Yale Univ. Press, 1923, doi: 10.1063/1.3061337.

[27] M. Zolfaghari, G. L. Oliveira, N. Sedaghat, and T. Brox, “Chained multi-stream networks exploiting pose, motion, and appearance for action classification and detection,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 2017, pp. 2904–2913, doi: 10.1109/ICCV.2017.316.

[28] P. Wang, W. Li, J. Wan, P. Ogunbona, and X. Liu, “Cooperative training of deep aggregation networks for RGB-D action recognition,” in Proc. AAAI Conf. Artif. Intell., vol. 32, no. 1, Apr. 2018, doi: 10.1609/aaai.v32i1.12228.

[29] F. Baradel, C. Wolf, and J. Mille, “Pose-conditioned spatio-temporal attention for human action recognition,” arXiv preprint arXiv:1703.10106, 2017, doi: 10.48550/arXiv.1703.10106.

[30] W. Ma, K. Wang, J. Li, S. X. Yang, J. Li, L. Song, and Q. Li, “Infrared and visible image fusion technology and application: A review,” Sensors, vol. 23, no. 2, p. 599, 2023, doi: 10.3390/s23020599.

[31] M. De Boissiere and R. Noumeir, “Infrared and 3D skeleton feature fusion for RGB-D action recognition,” IEEE Access, vol. 8, pp. 168297–168308, 2020, doi: 10.1109/ACCESS.2020.3023599.

[32] S. Hong, A. Ansari, G. Saavedra, and M. Martinez-Corral, “Full-parallax 3D display from stereo-hybrid 3D camera system,” Opt. Lasers Eng., vol. 103, pp. 46–54, 2018, doi: 10.1016/j.optlaseng.2017.11.010.

[33] G. Di Leo and A. Paolillo, “Uncertainty evaluation of camera model parameters,” in Proc. IEEE Int. Instrum. Meas. Technol. Conf. (I2MTC), May 2011, pp. 1–6, doi: 10.1109/IMTC.2011.5944307.

[34] M. Folk, G. Heber, Q. Koziol, E. Pourmal, and D. Robinson, “An overview of the HDF5 technology suite and its applications,” in Proc. EDBT/ICDT Workshop Array Databases, 2011, pp. 36–47, doi: 10.1145/1966895.1966900.

[35] S. Karim, G. Tong, J. Li, A. Qadir, U. Farooq, and Y. Yu, “Current advances and future perspectives of image fusion: A comprehensive review,” Inf. Fusion, vol. 90, pp. 185–217, 2023, doi: 10.1016/j.inffus.2022.09.019.

Applied and Basic Machine Intelligence Research

Vol. 3, No. 1, Spring & Summer 2025, pp. 35 - 50

A Novel Approach for Data Alignment in Infrared and RGB Image Fusion Algorithms

Raziyeh Razavi[1], Reza Rohani Sarvestani[2*]

[1]MSc., Department of Computer Engineering, Faculty of Technology and Engineering, Shahrekord University, Shahrekord, Iran

[2]Assistant Professor, Department of Computer Engineering, Faculty of Technology and Engineering, Shahrekord University, Shahrekord, Iran

Article Information

Original Research Paper

Received:
2025 February 13

Accepted:
2025 July 15

Keywords:
Image Fusion, Human Action Recognition, Data Alignment

Corresponding Author[*]:
rrohani@sku.ac.ir

Abstract

In recent years, human action recognition has become a key topic in the field of computer vision. However, one of the main challenges in this area is extracting effective features to enhance recognition accuracy. Infrared and RGB video data are commonly used for this purpose, yet neither of them alone provides a comprehensive representation of the scene. Therefore, combining these data types can lead to more accurate feature extraction. One effective approach to achieving this goal is through information fusion techniques. However, most human motion recognition datasets are not standardized for fusion, and the data are not properly aligned with each other.

In this study, the NTU RGB+D dataset is utilized, and a method for aligning and cropping video data is proposed to fuse infrared and RGB video frames. This method leverages inverse problem-solving techniques and body joint coordinates available in the dataset. The performance of the proposed approach is evaluated using EN, MI, SSIM, and MS-SSIM metrics. The obtained results indicate that the values of EN (7/17) and MI (13/1) demonstrate maximum information transfer and overlap. Additionally, the SSIM (0/78) and MS-SSIM (0/84) values confirm the preservation of structure and high quality of the fused data. These findings validate the effectiveness of the proposed method in enhancing video data fusion.

: 10.22034/ABMIR.2025.22795.1101

E-ISSN: 2821-2037

A Novel Approach for Data Alignment in Infrared and RGB Image Fusion Algorithms