Computational Intelligence in Electrical Engineering
Vol. 16, No. 3, 2025
pp. 99-120
Research Paper

A Framework for Video-Based Action Quality Assessment Using Sub-Score Learning and Temporal Dependencies

Marjan Mazruei[1], Ehsan Fazl-Ersi[2*], Abedin Vahedian[3], Ahad Harati[4]

[1] Department of Computer Engineering, Faculty of Engineering, Ferdowsi University of Mashhad, Mashhad, Iran

[2] Department of Computer Engineering, Faculty of Engineering, Ferdowsi University of Mashhad, Mashhad, Iran

[3] Department of Computer Engineering, Faculty of Engineering, Ferdowsi University of Mashhad, Mashhad, Iran

[3] Department of Computer Engineering, Faculty of Engineering, Ferdowsi University of Mashhad, Mashhad, Iran

Abstract:

Action Quality Assessment (AQA) is an emerging research problem aimed at providing automated, objective evaluation systems with broad applications across various domains. Most existing methods only predict the overall quality score and, due to the lack of fine-grained labels in datasets, are unable to evaluate the sub-stages of an action or provide detailed feedback. In this research, a novel framework for multi-stage action quality assessment is designed, focusing on model interpretability and the modeling of intra-action temporal dependencies in competitive sports videos. Employing a weakly supervised learning approach, the proposed method first divides videos into semantic clips comprising sub-actions and extracts their features. Then, by introducing a sequential pseudo-subscore learning mechanism, it models the causal dependencies among different stages of the movement. Unlike previous methods that assume that sub-actions are independent, the present framework incorporates the physical continuity of the movement into feature representations by injecting information from the preceding stage into the current one. Ultimately, a multi-stage score regression model is employed to predict subscores and provide greater model granularity. Experimental results on the UNLV Diving dataset demonstrate the strong and competitive performance of this method, achieving a Spearman's rank correlation coefficient of 0.90. Furthermore, statistical and visual analyses demonstrate that the subscores generated by the model are highly consistent with the judging criteria of the International Swimming Federation (FINA), effectively reflecting crucial stages such as the "Entry" in the final assessment.

Keywords: Action Quality Assessment, Temporal Semantic Segmentation, Pseudo-Subscore Generation, Multi-Substage Score Regression, Sequential Dependencies.

https://doi.org/10.22108/isee.2026.145975.1751

مقاله پژوهشی

چارچوبی برای ارزیابی کیفیت فعالیت در ویدئو مبتنی بر یادگیری زیرامتیازها و وابستگی‌های زمانی

مرجان مزروعی[۱]، احسان فضل ارثی[۲*]، عابدین واحدیان[۳]، احد هراتی[٤]

۱- گروه مهندسی کامپیوتر، دانشکده مهندسی، دانشگاه فردوسی مشهد، مشهد، ایران
marjan.mazruei@mail.um.ac.ir

۲- گروه مهندسی کامپیوتر، دانشکده مهندسی، دانشگاه فردوسی مشهد، مشهد، ایران
fazlersi@um.ac.ir

۳- گروه مهندسی کامپیوتر، دانشکده مهندسی، دانشگاه فردوسی مشهد، مشهد، ایران
vahedian@um.ac.ir

٤- گروه مهندسی کامپیوتر، دانشکده مهندسی، دانشگاه فردوسی مشهد، مشهد، ایران
a.harati@um.ac.ir

چکیده: ارزیابی کیفیت فعالیت یک مسئلهٔ پژوهشی نوظهور با هدف ارائهٔ سیستم‌های ارزیابی عینی خودکار است و کاربردهایی گسترده در حوزه‌های مختلف دارد. بیشتر روش‌های موجود فقط امتیاز کلی کیفیت را پیش‌بینی می‌کنند و به دلیل فقدان برچسب‌های ریزدانه در مجموعه‌داده‌ها، قادر به ارزیابی مراحل فرعی فعالیت و ارائهٔ بازخورد تفصیلی نیستند. در این پژوهش، یک چارچوب نوین برای ارزیابی چندمرحله‌ای کیفیت فعالیت با تمرکز بر تفسیرپذیری مدل و مدل‌سازی وابستگی‌های زمانی درونی فعالیت در ویدئوهای ورزشی رقابتی طراحی شده است. روش پیشنهادی با رویکرد یادگیری با نظارت ضعیف، ابتدا ویدئوها را به کلیپ‌های معنایی شامل زیرفعالیت‌ها تقسیم و ویژگی‌های آنها را استخراج می‌کند. سپس، با معرفی یک مکانیسم یادگیری شبه‌زیرامتیاز متوالی، وابستگی‌های علّی‌ومعلولی میان مراحل مختلف حرکت را مدل‌سازی می‌کند. برخلاف روش‌های پیشین که زیرفعالیت‌ها را مستقل فرض می‌کنند، چارچوب حاضر با تزریق اطلاعات مرحلهٔ قبل به مرحلهٔ فعلی، پیوستگی فیزیکی حرکت را در بازنمایی ویژگی‌ها لحاظ می‌کند. در نهایت، یک مدل رگرسیون امتیازی چندمرحله‌ای برای پیش‌بینی زیرامتیازها و ریزدانگی بیشتر مدل به کار گرفته می‌شود. نتایج آزمایش‌ها بر روی مجموعه‌دادهٔ UNLV-Diving نشان‌دهندهٔ عملکرد مطلوب و رقابتی این روش با دست‌یابی به ضریب همبستگی رتبهٔ اسپیرمن ۰/۹۰ است. همچنین، تحلیل‌های آماری و بصری اثبات می‌کنند زیرامتیازهای تولیدشده توسط مدل انطباق زیادی با معیارهای داوری فدراسیون جهانی شنا (FINA) دارند و به طور مؤثر مراحل حیاتی مانند «ورود به آب» را در ارزیابی نهایی بازتاب می‌دهند.

واژه‌های کلیدی: ارزیابی کیفیت فعالیت، بخش‌بندی معنایی زمانی، تولید شبه‌زیرامتیازهای زیرفعالیت‌ها، رگرسیون امتیازی چندمرحله‌ای، وابستگی‌های متوالی.

۱- مقدمه

با پیشرفت فناوری پردازش ویدئو و شبکه‌های عصبی عمیق، بسیاری از مطالعات بر مدل‌سازی رفتار انسان متمرکز شده‌اند. وظیفهٔ ارزیابی کیفیت فعالیت (AQA)[۱] مستلزم آن است که سیستمی هوشمند به طور خودکار و عینی[۲]، فعالیت انجام‌شده توسط انسان را از طریق ویدئوهای ورودی ارزیابی کند. برخلاف مسئلهٔ بازشناسی فعالیت انسان[۳] که نوع فعالیت را شناسایی می‌کند، در مسئلهٔ AQA، به جای

[۱] تاریخ ارسال مقاله : ۱٤۰٤/۰۲/۲٤
تاریخ پذیرش مقاله : ۱٤۰٤/۰۹/۰۸
نام نویسنده مسئول : احسان فضل ارثی

نشانی نویسنده مسئول : ایران، مشهد، دانشگاه فردوسی مشهد، دانشکده مهندسی

دسته‌بندی نمونه‌ها بر اساس نوع فعالیتشان، هدف اصلی این است که ارزیابی کند یک فعالیت چقدر خوب انجام شده است. سیستم‌های AQA دارای سناریوهای کاربردی عملی بسیاری از جمله پیش‌بینی امتیازهای رویدادهای ورزشی، مربی‌گری، رتبه‌بندی مهارت‌های جراحی، توان‌بخشی پزشکی و نظارت بر اقدامات خطرناک در تولیدات صنعتی هستند. مدل‌های AQA مبتنی بر یادگیری عمیق معمولاً از شبکه‌های عصبی کانولوشنی دوبُعدی یا سه‌بُعدی و LSTM برای استخراج و تجمیع ویژگی‌های ویدئوها و در نهایت تکمیل کار ارزیابی استفاده می‌کنند. روش‌های موجود برای ارزیابی کیفیت فعالیت عمدتاً بر اساس فرمول‌بندی مسئله به سه دستهٔ روش‌های مبتنی بر رگرسیون[۴]، مبتنی بر مقایسهٔ زوجی[۵] و مبتنی بر درجه‌بندی[۶] تقسیم می‌شوند.

در دستهٔ اول، **امتیازدهی رگرسیون**، هدف مدل پیش‌بینی یک امتیاز پیوسته برای هر ویدئو است و عموماً در حوزهٔ ورزش مشاهده می‌شود. از آنجا که داوران امتیازهای واقعی[۷] ویدئوها را در مجموعه‌داده‌های AQA قرار می‌دهند، مدل رگرسیون بردار پشتیبان (SVR)[۸] یا شبکهٔ کاملاً متصل (FCN) برای تکمیل پیش‌بینی امتیاز به طور مستقیم به کار گرفته می‌شود. میانگین مربعات خطا معمولاً به عنوان معیار متداول در امتیازدهی رگرسیون استفاده می‌شود. پژوهشگران در یک کار اولیه [۱]، با استفاده از تبدیل کسینوسی گسسته (DCT)[۹] و رگرسیون بردار پشتیبان، امتیازها را در ویدئوهای المپیک پیش‌بینی کردند. همچنین، پژوهش [۲] نشان داد آنتروپی تقریبی[۱۰] در ویژگی‌های حلت، اطلاعات حرکتی را بهتر از روش‌های DCT/DFT رمزگذاری می‌کند. لی و همکاران [۳] نیز یک روش یادگیری ویژگی حالت عمیق مبتنی بر اسکلت بدن ارائه دادند که در آن، یک شبکهٔ عصبی بازگشتی بر روی دنباله‌های اسکلت برای مدل‌سازی فعالیت‌های زمانی اعمال می‌شود.

تخمین حالت در فعالیت‌های ورزشی چالشی است، زیرا ویژگی‌های حالت معمولاً نشانه‌های بصری مانند پاشش آب[۱۱] در شیرجه را در نظر نمی‌گیرند. پارمار و موریس [٤] با بهره‌گیری از شبکهٔ کانولوشنی سه‌بُعدی C3D [٥] و LSTM، مدلی برای ارزیابی عملکرد بر پایهٔ نشانه‌های بصری ارائه دادند. لی و همکاران [٦] با ترکیب تابع هزینهٔ رتبه‌بندی[۱۲] و میانگین مربعات خطا، به دنبال در نظر گرفتن هم‌زمان دقت ارزش امتیاز و جنبه‌های رتبه‌بندی بودند. پارمار و موریس همچنین یک رویکرد چندوظیفه‌ای با زیرنویس ویدئویی و تشخیص عملکرد دقیق پیشنهاد کردند [۷]. آنها همچنین در [۸] مدل AQA را در یک رویکرد یادگیری چندوظیفه‌ای بررسی کردند که ترکیبی از بازشناسی دقیق فعالیت، تولید نظر و تفسیر عملکرد و پیش‌بینی امتیاز با استفاده از ویژگی‌های مکانی-زمانی حرکت و ظاهر و توابع هزینهٔ مناسب برای هر کدام از این وظایف است.

برای کاهش ضعف میانگین‌گیری یکسان ویژگی‌ها مانند [٤]، شیانگ و همکاران [۹] مدل S3D را بر اساس چارچوب بخش‌بندی زمانی نظارت‌شده ED-TCN[13] [۱۰] توسعه دادند. این رویکرد با تمرکز بر بخش‌های خاص فعالیت، نسبت به تحلیل کل ویدئو، عملکردی بهتر نشان داد. در [۱۱]، روشی ارائه شد که با پیش‌بینی ویژگی هر بخش، نمایش‌های مکانی-زمانی استخراج‌شده از شبکهٔ کانولوشنی سه‌بُعدی سنگین را به مدل‌های سبک‌تر دوبُعدی منتقل می‌کرد. به طور مشابه، [۱۲] با ترکیب شبکهٔ کانولوشنی سه‌بُعدی و Bi-LSTM، قطعات کلیدی را استخراج و از راهبرد تابع هزینهٔ دوگانه استفاده کرده است. در پژوهش [۱۳] نیز، پیش‌بینی امتیازهای زیرفعالیت‌ها[۱٤] با ترکیب بخش‌بندی معنایی زمانی و رگرسیون چندمرحله‌ای انجام شد. علاوه بر این، در [۱٤]، با بهره‌گیری از برچسب امتیاز کلی، شبه‌زیرامتیازهایی برای زیرفعالیت‌ها تولید شد؛ با این حال، نقطه ضعف اساسی این روش آن است که زیرفعالیت‌ها را به صورت کاملاً مستقل از یکدیگر ارزیابی و از پیوستگی فیزیکی و زنجیرهٔ علّی‌ومعلولی حرکات چشم‌پوشی می‌کند. مدل‌سازی عدم قطعیت باعث افزایش استحکام در امتیازدهی می‌شود. تلنگ و همکاران [۱٥] چارچوبی مبتنی بر توزیع امتیاز با در نظر گرفتن عدم قطعیت به منظور مقابله با ابهامات در امتیازدهی پیشنهاد دادند. از سوی دیگر، ژو و همکاران [۱٦] از خودرمزگذارهای واریاسیون شرطی[۱٥] برای مدل‌سازی عدم قطعیت ادراکی استفاده کردند.

در حوزهٔ توجه زمانی[۱٦]، ژو و همکاران [۱۷] مدل LSTM خودتوجهی[۱۷] را برای شناسایی حرکات فنی کلیدی

توسعه دادند. آنها همچنین از LSTM پرشی با هم‌پیوندی چندمقیاسی[۱۸] برای یادگیری وابستگی‌های زمانی محلی و جهانی بهره بردند. ژانگ و همکاران [۱۸] با ترکیب ویژگی‌های حرکت و حالت در یک معماری دوجریانی مبتنی بر شبکه‌های عصبی کانولوشنی گرافی[۱۹]، روابط بین قطعات ویدئویی را مدل کردند. در [۱۹]، مکانیسم توجه زمانی برای وزن‌دهی موثرتر به قطعات ویدئو نسبت به روش‌های تجمیعی[۲۰] معرفی شد. در ارزیابی مهارت‌های جراحی، وانگ و همکاران [۲۰] از یادگیری تقویتی[۲۱] برای بهینه‌سازی حالت‌ها استفاده کردند. فارابی و همکاران [۲۱] بر تجمیع وزن‌دار ویژگی‌ها تمرکز داشتند. همچنین، ژانگ و همکاران [۲۲] از مکانیسم‌های توجه زمانی بهره بردند. آنها همچنین در [۲۳] یک ماژول توجه غیرمحلی تفکیک‌پذیر[۲۲] را توسعه دادند که یکپارچگی ویژگی‌های اسکلتی و بصری را برای توالی‌های طولانی بهبود بخشید. آلالی و همکاران [۲٤] از ترنسفورمر مبتنی بر خودتوجهی خودکار برای ثبت وابستگی‌های بلندمدت استفاده کردند. ژو و همکاران [۲۵] توجه تقاطعی[۲۳] را برای استخراج ویژگی‌های آگاه از امتیاز به کار گرفتند.

رویکرد پیشنهادی در [۲٦] بر ویژگی‌های خاص ورزشکار با حذف پس‌زمینه تمرکز دارد. لی و همکاران [۲۷] نیز یک رمزگذار توالی فریم با راهنمایی گاوسی توسعه دادند. ژانگ و همکاران [۲۸] به منظور برطرف‌کردن ابهامات درون‌کلیپی و ناهمگونی بین‌کلیپی از شبکهٔ عصبی گراف کانولوشنی سلسله‌مراتبی بهره بردند. در [۲۹]، ماژول توجه مکان‌یابی چندمقیاسی برای شناسایی ورزشکاران در فریم‌ها به کار گرفته شد و ویژگی‌های استخراج‌شده توسط LSTMها برای ادغام اطلاعات توالی‌های محلی و جهانی استفاده شدند.

ژانگ و همکاران [۳۰] رمزگذار خودکار توزیعی برای کدگذاری توزیع امتیازها معرفی کردند. در ادامهٔ پژوهش آنها، [۳۱] با یکپارچه‌سازی توجه گروه‌آگاه و شبکه‌های عصبی گراف کانولوشنی، استخراج ویژگی‌های زمانی را بهبود داد. در [۳۲]، مدل IRIS برای پیش‌بینی امتیاز مبتنی بر روبریک ارائه شد. یک روش بلادرنگ [۳۳] نیز با بهره‌گیری از یک معماری سلسله‌مراتبی دولایه برای ویژگی‌های حالت و پویایی‌های زمانی-مکانی معرفی شد. روش‌های نیمه‌نظارت‌شده مانند [۳٤]، با بازیابی خودنظارتی ویژگی‌های بخش ماسک‌شده، از وابستگی‌های زمانی در ویدئوهای بدون برچسب بهره می‌گیرند و از ترکیب رگرسیون امتیاز و یادگیری تقابلی برای هم‌ترازسازی توزیع ویژگی‌ها میان داده‌های برچسب‌دار و بدون برچسب استفاده می‌کنند.

در حوزهٔ داده‌های چندوجهی، ادغام اطلاعات صوتی و بصری در ورزش‌هایی با موسیقی پس‌زمینه، دقت رگرسیون امتیاز را افزایش داده است. در [۳۵]، شبکه‌های فیوژن چندوجهی تطبیقی پیش‌رونده برای مدل‌سازی اطلاعات خاص حالت و ادغام ویژگی‌ها و تحلیل زمان‌بندی و ریتم فعالیت معرفی شده‌اند. داده‌های واحد اندازه‌گیری اینرسی و سیستم موقعیت‌یابی جهانی نیز می‌توانند به شبکه‌های عصبی عمیق در ارزیابی دقیق‌تر فعالیت‌ها کمک کنند [۳٦].

در دستهٔ دوم، یعنی **مقایسهٔ زوجی**، مدل دو ویدئو را از کتابخانهٔ ویدئویی به عنوان ورودی در نظر می‌گیرد. این دسته از روش‌ها کیفیت عملکرد دو ویدئو را به صورت مقایسه‌ای می‌سنجند و تلاش می‌کنند تا از طریق یادگیری فاصله یا امتیاز نسبی بین نمونه‌ها، مدل‌هایی بسازند که بتوانند رتبه‌بندی دقیق‌تری ارائه دهند. داتی و همکاران [۳۷] مدلی مبتنی بر رتبه‌بندی عمیق زوجی[۲٤] پیشنهاد دادند که از یک شبکهٔ عصبی کانولوشنی دومسیرهٔ سیامی[۲۵] بهره می‌برد. این مدل با دو تابع هزینهٔ رتبه‌بندی و شباهت آموزش داده شده است. در ادامه، آنها [۳۸] مدل توجه زمانی آگاه از رتبه را توسعه دادند که با استفاده از مکانیسم توجه قابل یادگیری و تابع هزینهٔ مبتنی بر رتبه، بر بخش‌های مرتبط با مهارت در ویدئو تمرکز می‌کند. لی و همکاران [۳۹] یک مدل توجه مکانی[۲٦] مبتنی بر شبکهٔ عصبی بازگشتی پیشنهاد کردند که در آن، از مکانیسم‌های ادغام توجه[۲۷] و تجمع زمانی[۲۸] برای بهبود استخراج ویژگی استفاده می‌شود. جین و همکاران [٤۰] از یک ماژول یادگیری عمیق متریک [٤۱] و بهره‌گیری از یک شبکهٔ سیامی برای یادگیری شباهت میان دو توالی فعالیت استفاده کرده‌اند. یو و همکاران [٤۲] نیز چارچوبی با عنوان رگرسیون تضادگونهٔ آگاه از گروه معرفی کردند که با ترکیب مقایسه‌های زوجی و درخت رگرسیون آگاه از گروه[۲۹]، به پیش‌بینی امتیاز عملکرد کمک می‌کند.

ژو و همکاران [٤۳] رویکردی آگاه از مراحل ارائه دادند

که با بهره‌گیری از ماژول توجه بخش‌بندی زمانی، توجه تقاطعی آگاه از رویه و رگرسیون متضاد در سطح زیرفعالیت، ارزیابی کیفی دقیق‌تری را ممکن کرده است. بای و همکاران [٤٤] نیز با استفاده از رمزگشایی مبتنی بر ترنسفورمر، نمایش‌های زمانی دقیق‌تری استخراج کردند و با به‌کارگیری توابع هزینهٔ نوآورانهٔ رتبه‌بندی و پراکندگی، عملکرد مدل را بهینه‌سازی کردند. به منظور یادگیری تفاوت‌های بین ویدئوها برای کمک به کار امتیازدهی نهایی، لی و همکاران [٤٥] یک شبکهٔ یادگیری متضاد زوجی طراحی کردند. در پژوهشی دیگر [٤٥]، یادگیری تضادگونه با رگرسیون ترکیب شد و از قید سازگاری برای حفظ انسجام میان خروجی‌ها استفاده شد. لیو و همکاران [٤٦] یک روش ارزیابی کیفیت فعالیت چندمرحله‌ای ارائه دادند که هدف آن بهبود دقت ارزیابی کیفیت فعالیت با در نظر گرفتن اطلاعات چندمرحله‌ای در ویدئوهای فعالیت است. اگرچه روش‌های مبتنی بر مقایسهٔ زوجی در مدیریت فرایند امتیازدهی مؤثر هستند، ایجاد مقایسه‌های زوجی معنادار نیازمند حجمی زیاد از داده‌های حاشیه‌نویسی‌شده است که محدودیتی مهم در مقیاس‌پذیری این روش‌ها محسوب می‌شود.

در دستهٔ سوم، **روش‌های مبتنی بر درجه‌بندی**، ویدئوها به چندین سطح عملکرد، برای مثال، مبتدی، متوسط و متخصص، تقسیم می‌شوند. شکل درجه‌بندی معمولاً در وظایف ارزیابی عملیات مهارت پزشکی ظاهر می‌شود. برای مثال، در [٤٧]، روشی مبتنی بر داده‌های سینماتیکی ربات همراه با ویژگی‌های استخراج‌شده مبتنی بر آنتروپی و ترکیب وزن‌دار ویژگی‌ها پیشنهاد شد که نسبت به روش‌های مبتنی بر مدل‌های مخفی مارکوف عملکردی بهتر ارائه می‌دهد. ضیاء و همکاران [٤٨] نیز با استفاده از معیارهای آنتروپی استخراج‌شده از داده‌های ویدئویی و شتاب‌سنج و با ترکیب فریم‌های ویدئو و حرکت اسکلتی، به بهبود آموزش مهارت‌های جراحی کمک کردند. این دسته از روش‌ها، زمانی که ارائهٔ بازخورد کیفی بدون نیاز به امتیازدهی عددی دقیق کافی باشد، عملکردی مناسب دارند. با این حال، به دلیل تکیه بر دسته‌بندی‌های گسسته و سخت‌گیرانه، معمولاً از درک تفاوت‌های ظریف در اجرای فعالیت‌ها بازمی‌مانند. این محدودیت، کاربرد آنها را در سناریوهای پیچیده و واقع‌گرایانه کاهش می‌دهد.

بیشتر روش‌های موجود یک امتیاز واحد برای کل ویدئو ارائه می‌دهند و ویژگی‌های سطح زیرفعالیت و وابستگی آنها را نادیده می‌گیرند؛ بنابراین، تفسیرپذیری کمی برای کاربر دارند. با این حال، مجموعه‌داده‌های موجود فاقد برچسب‌های صریح زیرامتیاز برای هر زیرفعالیت هستند. رسیدگی به این چالش نیازمند رویکردی مبتنی بر یادگیری با نظارت ضعیف است که بتواند وابستگی‌های زمانی بین زیرفعالیت‌ها را مدل‌سازی کند. این پژوهش یک چارچوب یادگیری شبه‌زیرامتیاز مبتنی بر رگرسیون امتیاز چندمرحله‌ای ارائه می‌دهد که شبه‌زیرامتیازها را به عنوان متغیرهای پنهان میانی با استفاده از استنتاج محلی تولید می‌کند. این چارچوب از بخش‌بندی معنایی زمانی برای تقسیم ویدئوها به زیرفعالیت‌های معنادار استفاده می‌کند. در این روش، ویژگی‌های مکانی–زمانی هر زیرفعالیت توسط شبکه‌های عصبی کانولوشنی سه‌بُعدی استخراج و با برچسب‌های امتیاز کلی به عنوان یک ویژگی کمکی برای تولید شبه‌زیرامتیازها برای داده‌های آموزشی تقویت می‌شوند. به منظور غلبه بر چالش استقلال زیرفعالیت‌ها، چارچوب پیشنهادی با بهره‌گیری از مفهوم «وابستگی متوالی محلی»، ویژگی‌های مکانی–زمانی را نه فقط با امتیاز کلی، بلکه با شبه‌زیرامتیاز تولیدشده در مرحلهٔ بلافاصله قبل غنی‌سازی می‌کند. این فرایند، بازنمایی هر زیرفعالیت را اصلاح می‌کند، جریان اطلاعات را در طول زمان حفظ و امکان تولید شبه‌زیرامتیازهای تفسیرپذیرتر و ارزیابی‌های دقیق کیفیت کلی فعالیت را فراهم می‌کند. در گام پایانی، آموزش تکمیلی مدل با حذف برچسب امتیاز کلی انجام می‌شود. مدل آموزش‌دیده زیرامتیازها و امتیاز کلی را با دقت زیاد پیش‌بینی می‌کند.

آزمایش‌ها بر روی مجموعه‌دادهٔ UNLV-Diving نشان‌دهندهٔ عملکرد مطلوب، رقابتی و بهبود قابل ملاحظهٔ این روش در مقایسه با روش‌های پایه هستند. این چارچوب وابستگی‌های زمانی زیرفعالیت‌های متوالی را مدل‌سازی و ضمن افزایش دقت ارزیابی، امکان ارائهٔ بازخوردی دقیق‌تر را برای شناسایی نقاط قوت و ضعف عملکرد بدون نیاز به حاشیه‌نویسی دستی گسترده فراهم می‌کند. همچنین، تحلیل‌های آماری نشان می‌دهند شبه‌زیرامتیازهای استخراج‌شده توسط این مدل انطباق زیادی با قوانین داوری

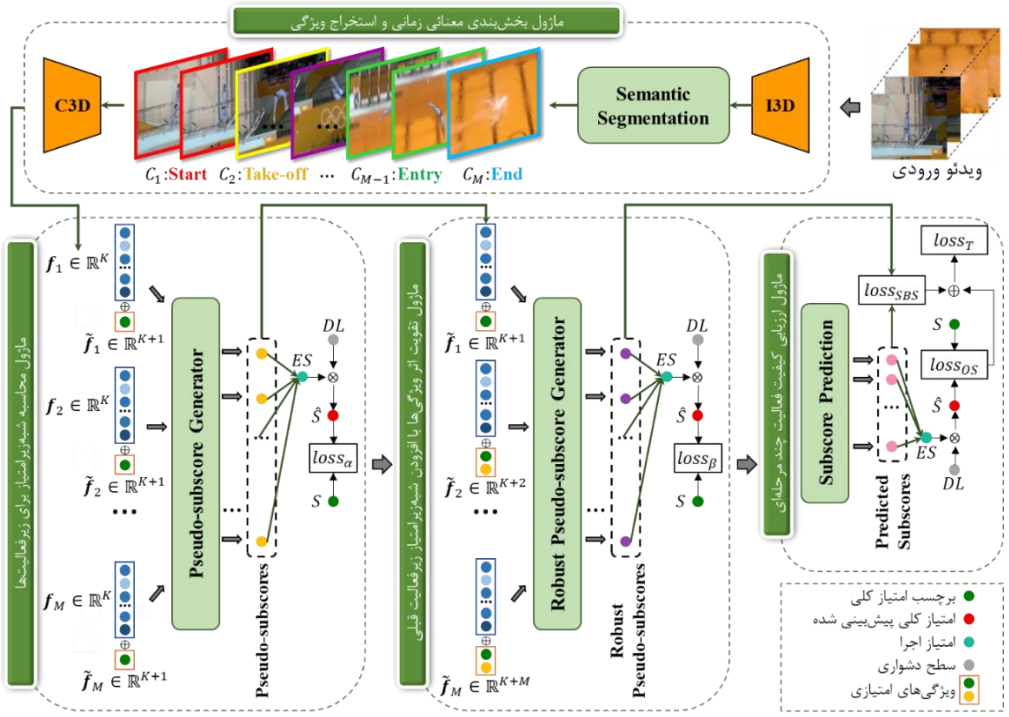
فدراسیون جهانی شنا (FINA) دارند که نشان‌دهندهٔ تفسیرپذیری زیاد این روش است. در ادامهٔ مقاله، بخش ۲ به شرح جزئیات چارچوب پیشنهادی اختصاص یافته است. نتایج تجربی در بخش ۳ بحث و بررسی شده‌اند و در نهایت، بخش ٤ به نتیجه‌گیری و جهت‌گیری‌های بالقوه برای پژوهش‌های آینده اشاره دارد.

۲- چارچوب پیشنهادی

در این پژوهش، رویکردی نوین برای ارزیابی کیفیت فعالیت در مراحل مختلف ویدئوهای ورزشی طراحی شده است. مسئلهٔ امتیازدهی در قالب یک چارچوب یادگیری با نظارت ضعیف و مبتنی بر رگرسیون بررسی شده است. چارچوب پیشنهادی با استفاده از بخش‌بندی معنایی زمانی فعالیت و در نظر گرفتن وابستگی‌های زمانی متوالی میان زیرفعالیت‌ها، به تحلیل نظام‌مند هر مرحله از اجرای ورزشکار اختصاص دارد؛ از این رو، با یادگیری شبه‌زیرامتیازها برای هر مرحله و تجمیع آنها، ارزیابی دقیق‌تری ارائه می‌دهد و امکان بازخورد تفصیلی دربارهٔ نقش هر زیرفعالیت در امتیاز کلی را فراهم می‌کند.

۲-۱- فرمول‌بندی مسئله

سیستم ارزیابی کیفیت فعالیت مبتنی بر ویدئو نوعی کار واکاوی فعالیت است که به طور خودکار، ارزیابی عینی کیفیت فعالیت خاص انسان را انجام می‌دهد. مجموعه‌دنباله‌های ویدئویی $V = \{\langle V_i, S_i\rangle | i = 1,2,\dots,N\}$ را در نظر بگیرید که شامل N ویدئو است و هر ویدئو به صورت $V_i = \{I_j\}$ نمایش داده می‌شود. ویدئوی $V_i \in \mathbb{R}^{T\times H\times W}$ شامل T فریم است و I_j نشان‌دهندهٔ فریم j–ام در ویدئوی V_i است. H و W به ترتیب ارتفاع و عرض هر فریم هستند. متغیر S_i نیز برچسب امتیاز کلی (واقعی) ویدئوی i–ام را نشان می‌دهد.



شکل (۱): مرور کلی چارچوب پیشنهادی. این چارچوب شامل چهار مرحلهٔ کلیدی است: (۱) ماژول بخش‌بندی معنایی زمانی و استخراج ویژگی: ویدئوی ورودی ابتدا به کلیپ‌هایی معنادار تقسیم می‌شود که هر کدام نشان‌دهندهٔ یک زیرفعالیت جداگانه هستند. سپس، ویژگی‌های مکانی–زمانی هر زیرفعالیت با استفاده از شبکه‌های عصبی کانولوشنی سه‌بُعدی از پیش آموزش‌دیده استخراج می‌شوند. (۲) ماژول محاسبهٔ شبه‌زیرامتیاز برای زیرفعالیت‌ها: شبه‌زیرامتیازها برای هر زیرفعالیت با ترکیب برچسب امتیاز کلی با ویژگی‌های استخراج‌شده تولید می‌شوند. (۳) ماژول تقویت اثر ویژگی‌ها با افزودن شبه‌زیرامتیاز زیرفعالیت قبلی به بردار ویژگی: وابستگی‌های زمانی متوالی با تقویت بازنمایی زیرفعالیت با بهره‌گیری از برچسب امتیاز کلی و شبه‌زیرامتیاز زیرفعالیت قبلی مدل‌سازی می‌شوند. (٤) ماژول ارزیابی کیفیت فعالیت چندمرحله‌ای: زیرامتیازهای نهایی با یکدیگر تجمیع می‌شوند تا امتیاز کلی با دقت زیاد پیش‌بینی شود.

هدف ارزیابی کیفیت فعالیت در این پژوهش پیش‌بینی دقیق امتیاز کلی عملکرد یک ورزشکار است. چارچوب پیشنهادی را می‌توان به صورت تابع نگاشت معادلهٔ (۱) فرمول‌بندی کرد:

$$\hat{S} = Q(V; w_{feat}, w_{tss}, w_{score}) \quad (۱)$$

که در آن، w_{feat}، w_{tss} و w_{score} به ترتیب پارامترهای سه ماژول پردازشی، یعنی ماژول استخراج ویژگی، ماژول بخش‌بندی معنایی زمانی و ماژول رگرسیون امتیازی را نشان می‌دهند. همچنین، $\hat{S}$ نشان‌دهندهٔ امتیاز پیش‌بینی‌شده است. همان‌طور که در شکل (۱) نشان داده شده است، چارچوب پیشنهادی از چهار بخش اصلی تشکیل شده است: مدل بخش‌بندی معنایی زمانی و استخراج ویژگی، مدل محاسبهٔ شبه‌زیرامتیازهای اولیه برای زیرفعالیت‌ها، مدل تقویت اثر ویژگی‌ها با افزودن شبه‌زیرامتیاز زیرفعالیت قبلی (مدل‌سازی وابستگی‌های زمانی متوالی) و مدل ارزیابی کیفیت فعالیت چندمرحله‌ای.

۲-۲- مدل بخش‌بندی معنایی زمانی و استخراج ویژگی

در این مرحله، برای به دست آوردن بازخورد دقیقِ کیفیت فعالیت و تفسیرپذیری بیشتر امتیاز کلی هر ورزشکار، ابتدا ویدئوی ورودی V_i به چندین کلیپ معنایی بر اساس مراحل فرعی[۳۰] یک فعالیت خاص در طول زمان تقسیم می‌شود؛ از این رو، تحلیلی دقیق‌تر از محتوای ویدئو امکان‌پذیر خواهد بود. از رویکرد پیشنهادی در [٤۳] به عنوان مدل برای بخش‌بندی معنایی زمانی استفاده می‌شود. به منظور شناسایی دقیق زیرفعالیت‌ها، ابتدا نقاط گذار زمانی[۳۱] بین مراحل مختلف فعالیت تشخیص داده می‌شوند. فرض کنید لازم است L نقطهٔ گذار زمانی شناسایی شوند. برای هر ویدئو، ماژول بخش‌بندی معنایی زمانی TSS احتمال وقوع مرحلهٔ گذار در فریم t را با استفاده از معادله‌های (۲) و (۳) پیش‌بینی می‌کند:

$$[\hat{p}_1, \dots, \hat{p}_L] = TSS(F(X)), \quad (۲)$$

$$\hat{t}_k = \underset{\frac{T}{L}(k-1)<t\leq\frac{T}{L}(k)}{\arg\max} \hat{p}_k(t) \quad (۳)$$

در اینجا، $\hat{p}_k(t)$ احتمال وقوع k–امین گذار زمانی در فریم t و $\hat{t}_k$ فریم پیش‌بینی‌شده برای آن گذار است. ماژول TSS از دو بلوک تشکیل شده است: بلوک پایین–بالا[۳۲] و بلوک خطی[۳۳]. بلوک پایین–بالا با بهره‌گیری از لایه‌های کانولوشن، طول ویژگی‌های I3D [٤۹] استخراج‌شده از ویدئو ($F(X)$) را در امتداد محور زمانی افزایش و به کمک لایه‌های حداکثر تجمیع[۳٤] در امتداد محور مکانی، ابعاد مکانی آن را کاهش می‌دهد. این فرایند موجب استخراج نمایش‌های عمیق‌تر مکانی–زمانی از ویژگی‌های بصری و زمینه‌ساز بخش‌بندی دقیق مراحل فعالیت می‌شود. خروجی بلوک پایین–بالا به بلوک خطی منتقل می‌شود تا توزیع‌های احتمال گذارهای زمانی برای L زیرفعالیت، یعنی مجموعهٔ $\{\hat{p}_k\}_{k=1}^{L}$، تولید شوند. برای تضمین ترتیب منطقی مراحل، شرط $\hat{t}_1 \leq \cdots \leq \hat{t}_L$ در معادلهٔ (۳) اعمال می‌شود.

برای آموزش، از توزیع واقعی گذار زمانی مرحلهٔ k–ام، یعنی t_k، یک توزیع دودویی p_k ساخته می‌شود که در آن، فقط $p_k(t_k) = 1$ و در سایر فریم‌ها $p_k(t_s)|_{s\neq k} = 0$ است. آنگاه، با استفاده از توزیع احتمال پیش‌بینی‌شده $\hat{p}_k$ و توزیع احتمال واقعی p_k، مسئلهٔ بخش‌بندی رویه به یک مسئلهٔ طبقه‌بندی متراکم تبدیل می‌شود که هدف آن پیش‌بینی احتمال تعلق هر فریم به گذار مرحلهٔ k–ام است. برای این منظور، از تابع هزینهٔ آنتروپی تقاطعی دودویی[۳۵] بین $\hat{p}_k$ و p_k برای بهینه‌سازی TSS مطابق معادلهٔ (٤) استفاده می‌شود:

$$loss_{BCE}(\hat{p}_k, p_k) = -\sum_t \big(p_k(t)\log\hat{p}_k(t) + (1 - p_k(t))\log(1-\hat{p}_k(t))\big) \quad (٤)$$

کمینه‌سازی $loss_{BCE}$ توزیع‌های $\hat{p}_k$ و p_k را به هم نزدیک‌تر می‌کند. پس از بخش‌بندی معنایی زمانی، هر کلیپ حاصل به یکی از زیرفعالیت‌های ویدئو مربوط است که باید ارزیابی شود. این فرایند به صورت آفلاین انجام می‌شود. در نهایت، ویدئوی V_i به مجموعه‌ای از کلیپ‌های ویدئویی معنایی $C = \{C_1, C_2, C_3, \dots, C_M\}$ تقسیم می‌شود که در

آن، $C_m \in \mathbb{R}^{T_m \times H \times W \times 3}$. H و W بـه ترتیـب عرض و ارتفاع تصـویر، T_m طول هر کلیپ و M تعداد کلیپ‌هاسـت. در نتیجه، بخش‌بندی معنایی زمانی می‌تواند مجموعه‌ای از کلیپ‌ها با طول‌های متفاوت، هر کدام شـامل یک زیرفعالیت مشـــخص، تولید کند. پس از بخش‌بندی معنایی زمـانی، مجموعه‌کلیپ‌های شـامل مراحل فرعی به ماژول اسـتخراج ویژگی ارسـال می‌شـوند. فرایند اسـتخراج ویژگی را می‌توان به صورت زیر نشان داد:

$$\mathbf{f}^{(ST)} = F_{DL}(C), \mathbf{f}_m^{(ST)} \in \mathbb{R}^K \qquad (۵)$$

که در آن، $F_{DL}(.)$ شـــبکهٔ عصـــبی عمیق اســـتخراج ویژگی‌های ویدئویی اسـت. اسـتخراج ویژگی یک ماژول باز اسـت که می‌تواند با بیشـتر شـبکه‌های یادگیری ویژگی، مانند شـبکه‌های عصـبی سـه‌بُعدی P3D [۵۰]، C3D [۵]، I3D [٤۹] و سـایر شـبکه‌های اسـتخراج ویژگی‌های ویدئویی جایگزین شـود. K نشـان‌دهندهٔ بُعد فضـای ویژگی اسـت. شبکه‌های عصبی کانولوشنی سه‌بُعدی، اگرچه به طور خاص برای تحلیل ویدئو طراحی نشـــده‌اند، به دلیل توانایی در مدل‌ســـازی هم‌زمان ویژگی‌های مکانی و زمانی، به یکی از روش‌های پرکاربرد در حوزهٔ تحلیل ویدئویی تبدیل شـده‌اند. در بسـیاری از پژوهش‌های اخیر، این شـبکه‌ها جایگزینی مناسـب برای مدل‌های سـنتی اسـتخراج ویژگی و بازنمایی ویدئوها محسـوب می‌شـوند. در این پژوهش نیز از شـبکهٔ C3D [۵] به عنوان شـبکهٔ اسـتخراج ویژگی اسـتفاده شـده اسـت. هنگام استفاده از C3D آموزش‌دیده، تعداد فریم‌ها در هر کلیپ روی یک مقدار ثابت تنظیم می‌شـود. دو دلیل رایج برای تعیین طول زمانی مشـــخص وجود دارنـد: محلیت و حافظه. نخســـت، برخلاف شـبکه‌های عصـبی کانولوشـنی دوبُعدی که صـــرفاً پیوسـتگی محلی مکانی را ثبت می‌کنند، شبکه‌های عصبی کانولوشنی سه‌بُعدی برای درک حرکت‌های محلی در بُعد زمانی طراحی شده‌اند. دوم، به دلیل محدودیت حافظه در پردازنده‌های گرافیکی، اسـتفاده از توالی‌های بلند در شـــبکه‌های عصـــبی کانولوشـــنی ســـه‌بُعدی به افزایش چشـــمگیر تعداد پارامترها و نیاز بیشـــتر به حافظه منجر می‌شـود. پس از اسـتخراج ویژگی، بردار ویژگی $\mathbf{f}_m^{(ST)}$ برای هر کلیپ شامل زیرفعالیت C_m به دست می‌آید.

۳-۲- مدل رگرسیون امتیازی

در این بخش، بـا اســـتفـاده از مجموعـه‌ویژگی‌هـای استخراج‌شده در سطح کلیپ‌های ویدئویی به عنوان ورودی، امتیازهای ارزیابی از طریق یک رگرســـیون چندمرحله‌ای به دسـت می‌آیند. به منظور حل مسـئلهٔ پیش‌بینی امتیاز برای هر زیرفعالیت با نظارت ضـــعیف، از یک شـــبکهٔ رگرســـیون جداگانه استفاده شـده است. به عبارت دیگر، یک زیرشـبکه برای هر زیرفعالیت طراحی شده است.

۱-۳-۲- محاسبهٔ شبه‌زیرامتیاز برای زیرفعالیت‌ها

به دلیل عدم وجود برچسب کیفیت برای هر زیرفعالیت، ارزیـابی دقیق عملکرد در هر مرحلـهٔ فرعی مســـئلـه‌ای چـالش‌برانگیز اســـت. همـهٔ مجموعـه‌داده‌هـای موجود فقط برچسـب امتیاز کلی را به عنوان شـاخص کیفیت فعالیت ارائه می‌کنند و فاقد حاشـــیه‌نویســـی امتیاز برای هر زیرفعالیت هسـتند. علاوه بر این، حاشـیه‌نویسـی دسـتی امتیازهای هر زیرفعالیت به دانش و تجربهٔ حرفه‌ای نیاز دارد که به‌ســـختی به دســـت می‌آید. برای برطرف‌کردن این محدودیت، مدلی پیشـنهاد شـده اسـت که با پیش‌بینی شـبه‌زیرامتیازها برای زیرفعالیت‌ها، امتیاز کلی را از طریق ادغام این مقادیر محاسبه می‌کند. در **گام ابتدایی** رگرسـیون امتیاز پیشـنهادی، برای هر زیرفعالیت یک مقدار شـــبه‌زیرامتیاز تولید می‌شـــود. پس از اســـتخراج ویژگی‌های مکانی-زمانی، کلیپ‌های ویدئویی مربوط به زیرفعالیت‌ها به مجموعه‌ای از بردارهای ویژگی به صـورت $\mathbf{f}^{(ST)} = \left\{\mathbf{f}_1^{(ST)}, \dots, \mathbf{f}_M^{(ST)}\right\}$ تبدیل می‌شـوند که در آن، هر بردار $\mathbf{f}_m^{(ST)} \in \mathbb{R}^K$ نمایانگر زیرفعالیت m-ام اسـت. در این گام، غنی‌سازی ویژگی‌ها با ساخت ویژگی‌های جدید $(\tilde{\mathbf{f}})$ از طریق ترکیـب وزن‌دار ویژگی‌هـای مکـانی-زمـانی $(\mathbf{f}^{(ST)})$ و ویژگی‌هـای امتیـازدهی $(\mathbf{f}^{(SS)})$ انجام می‌شـــود. ویژگی‌های مکانی-زمانی، ویژگی‌های بصـری و سـاختاری ورودی را ثبت می‌کنند؛ در حالی که ویژگی‌های امتیازدهی،

اطلاعات ارزیابی زمینه‌ای را ارائه می‌دهند. بردار ویژگی تقویت‌شده برای هر زیرفعالیت با استفاده از معادلهٔ (٦) به دست می‌آید:

$$\tilde{\mathrm{f}}_m = \alpha \times \mathrm{f}_m^{(ST)} \oplus (1-\alpha) \times \mathrm{f}_m^{(SS)} \quad (٦)$$

مشابه رویکرد پیشنهادی در پژوهش [١٤]، در این مرحله، برچسب امتیاز کلی به عنوان یک ویژگی امتیازدهی به بردار ویژگی اضافه می‌شود. به عبارتی، در معادلهٔ (٦)، $\mathrm{f}_m^{(SS)} = [S]$ است. پارامتر $\alpha \in [0,1]$ به عنوان ضریب وزنی، تعادلی میان سهم امتیاز کلی و ویژگی‌های اصلی برقرار می‌کند. نماد $\oplus$ نیز بیانگر عملیات الحاق ویژگی‌هاست. این عملیات به تولید مجموعه‌ویژگی جدید $\{\tilde{\mathrm{f}}_1, \dots, \tilde{\mathrm{f}}_M\}$ منجر می‌شود که در آن، $\tilde{\mathrm{f}}_m \in \mathbb{R}^{(K+1)}$ است. این مجموعه‌ویژگی‌های بهبودیافته متعاقباً به یک شبکهٔ کاملاً متصل با پنج لایه وارد می‌شوند. لایهٔ اولیه بردارهای ویژگی بهبودیافته را به عنوان ورودی دریافت می‌کند و تعداد گره‌ها در هر لایه به‌تدریج به ۱ کاهش می‌یابد. در لایهٔ نهایی، از تابع فعال‌سازی سیگموئید برای پیش‌بینی شبه‌زیرامتیاز هر زیرفعالیت استفاده می‌شود. همهٔ زیرشبکه‌ها از ساختاری یکسان استفاده می‌کنند. فرایند محاسبهٔ شبه‌زیرامتیاز را می‌توان به صورت زیر توصیف کرد:

$$\begin{aligned} \tilde{\mathrm{f}}_m^t &= FC\left(W^t, \tilde{\mathrm{f}}_m^{(t-1)}\right) \\ sbs_m &= Sigmoid(\mathrm{f}'_m) \end{aligned} \quad (٧)$$

در اینجا، $FC(\cdot)$ شبکهٔ کاملاً متصل، f'_m خروجی آخرین لایه، W^t نشان‌دهندهٔ پارامترهای لایهٔ t–ام و sbs_m شبه‌زیرامتیاز پیش‌بینی‌شده برای زیرفعالیت m–ام است. این فرایند تولید شبه‌زیرامتیاز برای تمام زیرفعالیت‌ها تکرار می‌شود تا شبه‌زیرامتیازها به طور مستقل پیش‌بینی شوند. پس از تولید شبه‌زیرامتیازها برای همهٔ زیرفعالیت‌ها، آنها به یک شبکهٔ کاملاً متصل وارد می‌شوند تا امتیاز کلی ویدئو پیش‌بینی شود. این فرایند به صورت زیر بیان می‌شود:

$$\hat{S} = sigmoid\left(\sum_{m=1}^{M}(w_m \times sbs_m)\right), \hat{S} \in \mathbb{R} \quad (٨)$$

در این معادله، $\hat{S}$ نشان‌دهندهٔ امتیاز کلی پیش‌بینی‌شده است و w_m به پارامترهای وزنی لایهٔ کاملاً متصل اشاره دارد. برای آموزش این شبکه، میانگین مربعات خطا به عنوان تابع هزینه بین امتیاز پیش‌بینی‌شده و برچسب امتیاز کلی هر ویدئو اعمال می‌شود. تابع هزینه را می‌توان به شکل معادلهٔ (٩) فرمول‌بندی کرد:

$$loss_\alpha = \frac{1}{N_{Tr}}\sum_{i=1}^{N_{Tr}}\left(\hat{S}_i - S_i\right)^2 \quad (٩)$$

در اینجا، N_{Tr} تعداد ویدئوهای آموزشی است. $\hat{S}_i$ امتیاز کلی پیش‌بینی‌شده برای نمونهٔ آموزشی i–ام و S_i برچسب امتیاز کلی مربوط است. مدل آموزش‌دیده ویدئوهای آموزشی را پردازش می‌کند تا شبه‌زیرامتیازهایی را برای زیرفعالیت‌های هر ویدئو استخراج کند که با $SBS = \{sbs_1, \dots, sbs_M\}$ نشان داده می‌شوند. این رویکرد جامع تضمین می‌کند مدل از داده‌های آموزشی به طور مؤثر برای تولید شبه‌زیرامتیازها استفاده می‌کند.

۲-۳-۲- تقویت اثر ویژگی‌ها با افزودن شبه‌زیرامتیاز زیرفعالیت قبلی به بردار ویژگی (مدل‌سازی وابستگی‌های زمانی متوالی)

افزودن صِرف برچسب امتیاز کلی به عنوان یک ویژگی برای تمام زیرفعالیت‌ها به ایجاد نوعی سوگیری در فرایند یادگیری منجر می‌شود. این موضوع باعث می‌شود زیرامتیازهای پیش‌بینی‌شده به‌طور غیرواقعی با امتیاز کلی هم‌راستا شوند، که در نتیجه، تنوع آنها کاهش می‌یابد و توان مدل در تشخیص تغییرات موضعی در طول فعالیت محدود می‌شود. به طور خاص، این بخش وابستگی‌های متوالی بین زیرمرحله‌های درون یک فعالیت را بررسی می‌کند. درک این وابستگی‌ها برای طراحی یک معماری مؤثر که امتیاز کیفیت فعالیت را به طور دقیق تخمین می‌زند، بسیار مهم است.

روابط زمانی میان زیرفعالیت‌های متوالی در ارزیابی کیفیت فعالیت نقش کلیدی دارند؛ زیرا عملکرد هر زیرفعالیت ممکن است مستقیماً بر زیرفعالیت بعدی تأثیرگذار باشد. برای مثال، در فرایند شیرجه، یک خطای کوچک در مرحلهٔ پرش[٣٦] ممکن است روند چرخش[٣٧] یا

پشـــتک‌زدن[۳۸] را مختل کند و در نهایت باعث افت کیفیت اجرای کلی شود. به طور مشابه، فرم نامناسب بدن یا سرعت ناکافی در هنگام پشتک ممکن است چرخش را ضعیف کند و بر کیفیت ورود به آب اثر منفی بگذارد. مرحلهٔ ورود به آب یکی از اجزای حیاتی در ارزیابی شـــیرجه محســـوب می‌شـود. ورود صـحیح با بدن کشـیده و بدون زاویه پاشـش آب را به حداقل می‌رساند و نشان‌دهندهٔ دقت و کنترل زیاد ورزشـــکار اســـت. در مقابل، ورود ناهماهنگ و عدم حفظ موقعیت مناسب بدن معمولاً با پاشش شدید آب همراه است و از کیفیت کلی شیرجه می‌کاهد.

پس از تولید شـــبه‌زیرامتیازها، این مقادیر به صـــورت متوالی به عنوان ویژگی‌های اضـافی برای تقویت بیشـتر مدل به‌کار گرفته می‌شـوند. این فرایند، با ادغام اطلاعات زمینه‌ای غنی‌تر در نمـایش ویژگی‌هـا، موجـب توانمنـدی مـدل برای محاســـبهٔ شـــبه‌زیرامتیازهای دقیق‌تر و صـــحیح‌تر برای هر زیرفعالیت می‌شـــود. **گام دوم** چارچوب رگرســـیون امتیاز پیشـــنهادی برای تولید مقادیر شـــبه‌زیرامتیاز قوی‌تر برای هر زیرفعالیت طراحی شده است. در این مرحله، برچسب امتیاز کلی و شـــبه‌زیرامتیاز زیرفعالیت قبلی به عنوان ویژگی‌های اضـافی در نظر گرفته می‌شـوند. مشـابه بخش قبلی، تقویت ویژگی‌ها از طریق ترکیب وزن‌دار ویژگی‌های مکانی–زمانی و ویژگی‌هـای امتیـازدهی انجـام می‌شـــود. در این حـالـت، ویژگی‌هـای امتیـازدهی از هر دو برچســـب امتیـاز کلی و شـــبه‌زیرامتیاز زیرفعالیت بلافاصـــله قبل از آن اســـتخراج می‌شـــوند و اطلاعات ارزیابی زمینه‌ای را فراهم می‌کنند. به طور خـاص، بردار ویژگی‌هـای امتیـازدهی در معـادلـهٔ (٦) معادل $\mathbf{f}_m^{(SS)} = [S, sbs_{m-1}]$ است.

ویژگی‌هـای تقویـت‌شـــده در ادامـه برای محاســـبهٔ شـــبه‌زیرامتیازهای قوی‌تر اســـتفاده می‌شـــوند. مراحل بعدی آموزش مدل شـامل تولید شـبه‌زیرامتیازها و پیش‌بینی امتیاز کلی، طبق رویهٔ ارائه‌شـده در زیربخش قبلی انجام می‌شـوند. پس از آموزش کامل، شـبه‌زیرامتیازهای به‌روزرسـانی‌شـده برای زیرفعـالیـت‌هـای هر ویـدئوی آموزشـــی از مـدل آموزش‌دیده اســـتخراج می‌شـــوند که با نماد $SBS^{new} = \{sbs_1, ..., sbs_M\}$، نمایش داده می‌شوند.

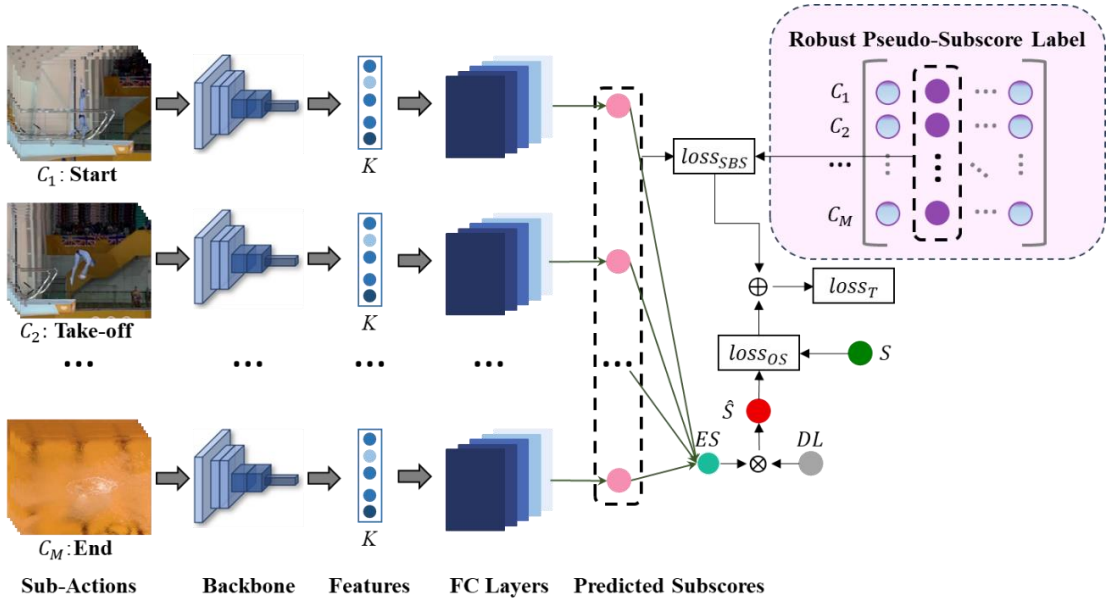
۳-۳-۲- مدل ارزیابی کیفیت فعالیت چندمرحله‌ای

در نهایت، پس از مدل‌سـازی وابسـتگی‌های زمانی میان زیرفعالیت‌ها و محاســـبهٔ شـــبه‌زیرامتیازهای قوی برای هر زیرفعالیت، هدف دســـت‌یابی به امتیاز نهایی با دقتی زیاد است. از آنجا که در فاز استنتاج (آزمون) برچسب امتیاز کلی نامشخص است و باید پیش‌بینی شود، این برچسب از بردار ویژگی ورودی حذف می‌شـود و یک فرایند آموزش تکمیلی (Fine-tuning) صـــرفاً بر پایهٔ ویژگی‌های بصـــری (بدون ویژگی برچسب امتیاز کلی) انجام می‌شود.

در این مرحله، خروجی شـبکهٔ اسـتخراج ویژگی به یک شبکهٔ کاملاً متصل با پنج لایهٔ متوالی وارد می‌شود. لایهٔ اولیهٔ این شبکه بردار ویژگی K–بُعدی را به عنوان ورودی دریافت می‌کند و با کاهش تدریجی تعداد گره‌ها در هر لایه، در نهایت آن را به یک گره منفرد می‌رساند. نحوهٔ محاسبهٔ این فرایند در معادلات (۱۰) شرح داده شده است:

$$\begin{aligned} \mathbf{f}_m^t &= FC\left(W^t, \mathbf{f}_m^{(t-1)}\right) \\ S_m^{sub} &= sigmoid(\mathbf{f'}_m) \\ \hat{S} &= sigmoid\left(\sum_{m=1}^{M} w_m \times S_m^{sub}\right) \end{aligned} \qquad (۱۰)$$

در اینجا، S_m^{sub} به عنوان زیرامتیاز مربوط به زیرفعالیت m–ام تعریف می‌شـــود. مطابق توضـــیحات زیربخش قبلی، برای هر ویدئوی آموزشـی، برچسـب‌های شـبه‌زیرامتیاز قوی به دست می‌آیند. تابع هزینهٔ استفاده‌شده در آموزش این مدل از دو بخش تشـکیل شـده اسـت. بخش اول میانگین مربعات خطا بین زیرامتیازهای پیش‌بینی‌شـده و شـبه‌زیرامتیازها برای هر زیرفعـالیـت را محاســـبـه می‌کنـد؛ و بخش دوم میـانگین مربعات خطا بین امتیاز کلی پیش‌بینی‌شـده و برچسـب امتیاز کلی را ارزیابی می‌کند. تابع هزینهٔ نهایی مدل از ترکیب دو مؤلفهٔ یادشـده سـاخته می‌شـود (همان‌طور که در شـکل (۲) نشان داده شده است):



شکل (۲): مدل ارزیابی کیفیت فعالیت چندمرحله‌ای.

$$loss_m^{SBS} = (S_m^{sub} - sbs_m)^2$$
$$loss_{OS} = (\hat{S} - S)^2 \quad (۱۱)$$
$$loss_T = loss_{OS} + \sum_{m=1}^{M} loss_m^{SBS}$$

پس از پایان آموزش تکمیلی، ویدئوهای مجموعه‌تست به مدل آموزش‌دیده تغذیه می‌شوند و در نتیجه، زیرامتیازهای مربوط به زیرفعالیت‌های مختلف هر ویدئو و همچنین امتیاز کلی آن با دقت زیاد پیش‌بینی می‌شوند.

۳- ارزیابی کارایی چارچوب پیشنهادی

۳-۱- مجموعه‌داده و معیارهای ارزیابی

مجموعه‌داده: در این مطالعه، مجموعه‌دادهٔ UNLV-Diving [۴] برای ارزیابی عملکرد و اعتبارسنجی اثربخشی چارچوب پیشنهادی استفاده شده است. مجموعه‌دادهٔ UNLV-Diving گسترش‌یافتهٔ مجموعه‌دادهٔ MIT-Diving [۱] است. این مجموعه‌داده شامل ۳۷۰ نمونهٔ ویدئویی از اجرای شیرجه‌های انفرادی است که از مرحلهٔ نیمه‌نهایی و فینال رقابت سکوی ۱۰ متر مردان در المپیک تابستانی ۲۰۱۲ لندن گردآوری شده‌اند. هر ویدئو با مدت زمان تقریبی ۴ ثانیه، از پلتفرم یوتیوب استخراج و با امتیاز کلی، سطح دشواری و برچسب‌های بخش‌بندی زیرفعالیت در طول زمان حاشیه‌نویسی شده است که امکان محاسبهٔ امتیازهای اجرا را فراهم می‌کند. سطوح دشواری اجراها در بازه‌ای بین ۲/۷ تا ۴/۱ و امتیاز کلی در محدودهٔ ۲۱/۶ تا ۱۰۲/۶ متغیر است. تمامی فریم‌های ویدئویی با وضوح ۳۲۰×۲۴۰ پیکسل دخیره شده‌اند. در میان تمامی ویدئوها، ۳۰۰ ویدئو برای آموزش و ۷۰ ویدئو برای تست استفاده شده‌اند.

معیارهای ارزیابی: در بیشتر مطالعات، از ضریب همبستگی رتبهٔ اسپیرمن ρ (SRC) به عنوان معیار ارزیابی اصلی استفاده می‌شود. ضریب همبستگی رتبهٔ اسپیرمن، یا به‌اختصار همبستگی اسپیرمن، یک معیار آماری غیرپارامتری است که توسط چارلز اسپیرمن در سال ۱۹۰۴ پیشنهاد شد و این نام را از آن گرفته است. SRC می‌تواند همبستگی بین دو متغیر آماری را ارزیابی کند. مقدار همبستگی اسپیرمن بین ۱- و ۱ قرار دارد. هرچه به ۱ نزدیک‌تر باشد، همبستگی بیشتر است. اگر ۱ باشد، به این معناست که دو متغیر تصادفی کاملاً همبستگی مثبت دارند، ۰ به معنای عدم همبستگی و نامرتبط‌بودن آنها و ۱- به این معناست که رابطهٔ بین آنها یک همبستگی کاملاً منفی است. همبستگی اسپیرمن معیاری از رابطهٔ یکنواخت بین دو متغیر تصادفی است که به طور هم‌زمان تغییر می‌کنند، اما نه با سرعت یکسان؛ بنابراین، رابطه لزوماً خطی نیست. در اینجا، SRC توافق بین رتبه‌بندی‌های پیش‌بینی‌شده و رتبه‌بندی‌های واقعی ورزشکاران را ارزیابی می‌کند. ضریب همبستگی رتبهٔ

اسپیرمن به صورت زیر تعریف می‌شود:

$$\rho = \frac{cov(p,q)}{\sigma_p \sigma_q} = \frac{\sum_i (p_i - \bar{p})(q_i - \bar{q})}{\sqrt{\sum_i (p_i - \bar{p})^2 \sum_i (q_i - \bar{q})^2}} \quad (۱۲)$$

که در آن، p و q به ترتیب دنبالهٔ رتبه‌بندی امتیازهای پیش‌بینی‌شده و امتیازهای واقعی هستند. cov کوواریانس این دو دنبالهٔ رتبه‌بندی است و σ_p و σ_q انحرافات معیار استاندارد p و q هستند. هرچه مقدار ρ بزرگ‌تر باشد، مقدار پیش‌بینی‌شده به مقدار واقعی نزدیک‌تر است و دقت روش بیشتر است. در وظیفهٔ AQA، شاخص ارزیابی کلی و عمومی SRC است، اما روش‌هایی نیز وجود دارند که از MSE و MED برای ارزیابی عملکرد استفاده می‌کنند (معادله‌های (۱۳) و (۱٤)). به ترتیب، MSE میانگین مربعات اختلاف بین امتیازهای پیش‌بینی‌شده و واقعی را اندازه‌گیری می‌کند؛ به طور مشابه، MED میانگین اختلاف مطلق بین امتیازهای پیش‌بینی‌شده و واقعی را اندازه‌گیری می‌کند:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}\left(S_i - \hat{S}_i\right)^2 \quad (۱۳)$$

$$MED = \frac{1}{N}\sum_{i=1}^{N}\left|S_i - \hat{S}_i\right| \quad (۱٤)$$

$\hat{S}_i$ نشان‌دهندهٔ مقدار امتیاز پیش‌بینی‌شده و S_i نشان‌دهندهٔ مقدار امتیاز واقعی و N تعداد کل نمونه‌هاست. هرچه هر یک از این مقادیر معیار کوچک‌تر باشد، فاصلهٔ بین امتیاز پیش‌بینی‌شده و مقدار واقعی کمتر است؛ بنابراین، عملکردی بهتر خواهد داشت. در این کار، از تمام این معیارهای ارزیابی – SRC، MSE و MED – برای اندازه‌گیری دقت امتیاز پیش‌بینی‌شده استفاده می‌شود.

۲-۳- جزئیات اجرا

چارچوب پیشنهادی برای ارزیابی عملکرد ورزشکاران از طریق یک رویکرد امتیازدهی چندمرحله‌ای، مطابق معیارهای ارزیابی عینی ورزشی، طراحی شده است. در برپایش تجربی[۳۹]، یک مدل بخش‌بندی معنایی زمانی، مبتنی بر چارچوب [٤۳]، برای تقسیم هر ویدئو به زیرفعالیت‌های مجزا به کار گرفته می‌شود. فرایند شیرجه به پنج زیرفعالیت خاص – شروع، پرش، چرخش[٤۰]، ورود به آب و پایان – تقسیم می‌شود تا حرکات دقیق ثبت شوند. برای بازنمایی ویژگی‌ها، از یک مدل شبکهٔ عصبی کانولوشنی سه‌بُعدی C3D [٥] از پیش آموزش‌دیده استفاده می‌شود. ویژگی‌های استخراج‌شده توسط این مدل با ابعاد ٤۰۹٦ مشخص می‌شوند که بازنمایی قوی برای هر زیرفعالیت را تضمین می‌کند. در طول فرایند آموزش و تست، ۱٦ فریم از هر زیرفعالیت انتخاب و به ۱۱۲×۱۱۲ پیکسل تغییر اندازه داده می‌شوند. فرایند آموزش از دو نوع برچسب استفاده می‌کند: امتیاز کلی و امتیاز اجرا برای هر فعالیت. امتیاز کلی برای عملکرد شیرجه به عنوان حاصل‌ضرب امتیاز اجرا و سطح دشواری فعالیت محاسبه می‌شود. برای حفظ سازگاری، امتیازات کلی با استفاده از نرمال‌سازی حداقل–حداکثر نرمال‌سازی می‌شوند، در حالی که امتیازات اجرا که از ۰ تا ۳۰ متغیر هستند، با تقسیم بر ۳۰ نرمال‌سازی می‌شوند.

چارچوب پیشنهادی بر روی یک پردازندهٔ گرافیکی NVIDIA RTX 2080 با ۱۱ گیگابایت حافظهٔ داخلی اجرا شده است. زبان برنامه‌نویسی پایتون ۳ بوده و برای اجرای شبکه‌های عصبی عمیق از کتابخانهٔ پایتورچ [٥۱] در سیستم عامل لینوکس استفاده شده است. ترکیبی از شبکه‌های کاملاً متصل که به طور خاص برای هر زیرفعالیت طراحی شده‌اند، برای آموزش استفاده می‌شود. تابع هزینهٔ MSE برای بهینه‌سازی امتیازهای پیش‌بینی‌شده استفاده می‌شود. برای کاهش بیش‌برازش، تنظیم[٤۱] L2 با پارامتر کاهش وزن[٤۲] ۰/۰۰۰٥ اعمال می‌شود و احتمال رهاکردن[٤۳] ۰/٥ پس از لایه‌های تجمیع متوسط[٤٤] گنجانده می‌شود. فرایند آموزش با استفاده از بهینه‌ساز تخمین گشتاور وفقی (Adam)[٤٥] با نرخ یادگیری[٤٦] اولیهٔ ۰/۰۰۰۱ انجام می‌شود که هر ۳۰ تکرار آموزش با ضریب ۰/۱ کاهش می‌یابد. این مطالعه همچنین پیکربندی‌های مختلف شبکه را بررسی کرد و دریافت معماری پنج لایه بهترین تعادل بین دقت و هزینهٔ محاسباتی را فراهم می‌کند. نتایج نهایی با میانگین‌گیری معیارهای SRC، MSE و MED در ۱۰ تکرار آزمایش به دست آمد.

۳-۳- مقایسه با روش‌های ارزیابی کیفیت فعالیت موجود

به منظور ارزیابی عملکرد چارچوب پیشنهادی، نتایج

حاصل از اجرای آن با نتایج چندین روش موجود، با استفاده از مجموعه‌دادهٔ UNLV-Diving مقایسه شده است. خلاصهٔ این مقایسه در جدول (۱) ارائه شده است. اعداد ارائه‌شده برای روش‌های مرجع مستقیماً از مقاله‌های اصلی استخراج شده‌اند. گفتنی است، مقادیر مشخص‌شده با علامت «–» در جدول به این دلیل وارد نشده‌اند که نتایج مربوط به آن معیارها در مقالهٔ اصلی مربوط گزارش نشده‌اند. همان‌طور که از نتایج جدول مشخص است، روش پیشنهادی در تمامی معیارهای ارزیابی عملکردی بسیار مطلوب و رقابتی نسبت به بیشتر روش‌های موجود از خود نشان داده است. به طور خاص، این روش با دست‌یابی به ضریب همبستگی رتبهٔ اسپیرمن ۰/۹۰، همبستگی قوی بین رتبه‌بندی امتیازهای پیش‌بینی‌شده و واقعی را اثبات می‌کند. همچنین، با کاهش مقادیر MSE و MED به ترتیب تا ۳۲/۴۲ و ۲/۸۳، میزان خطای پیش‌بینی کاهش یافته است.

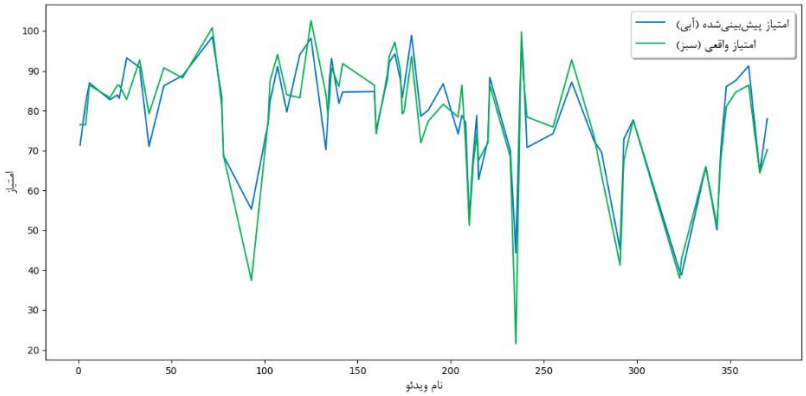
همان‌طور که در جدول (۱) مشاهده می‌شود، چارچوب پیشنهادی با سایر روش‌های نوین ارزیابی کیفیت فعالیت از جمله مرجع [۱۴] و رویکرد اخیر ارائه‌شده در پژوهش [۵۲] مقایسه شده است. نتایج نشان می‌دهد مدل‌ها از نظر نحوهٔ رویارویی با وابستگی‌های زمانی، رفتارها و دقت‌هایی متفاوت از خود نشان می‌دهند. در مقایسه با مرجع [۱۴]، مدل پیشنهادی توانسته است مقدار SRC را به میزان ۰/۱ افزایش و مقادیر MSE و MED را به ترتیب ۵۳/۹۵ و ۴/۷۶ واحد کاهش دهد. دلیل این برتری آن است که مرجع [۱۴] برای جبران کمبود برچسب فقط به ترکیب امتیاز کلی با ویژگی‌ها بسنده می‌کند و توالی منطقی حرکات را نادیده می‌گیرد (استقلال زیرفعالیت‌ها). در مقابل، چارچوب پیشنهادی ما با تزریق شبه‌زیرامتیاز مرحلهٔ پیشین (sbs_{m-1}) توانسته است یک بافت زمانی محلی[۴۷] ایجاد و تأثیر زنجیره‌وار کیفیت زیرفعالیت‌ها بر یکدیگر را به‌خوبی مدل‌سازی کند.

از سوی دیگر، با مقایسهٔ نتایج روش پیشنهادی با پژوهش [۵۲]، مشاهده می‌شود مدل ارائه‌شده در [۵۲] به نتایج کمّی بهتری دست یافته است. این اختلاف در نتایج ریشه در دو تفاوت ساختاری دارد: نخست آنکه پژوهش [۵۲] به جای وابستگی محلی، از یک تاریخچهٔ تجمعی بلندمدت[۴۸] بهره می‌برد که اطلاعات تمامی مراحل پیشین را تجمیع می‌کند تا تأثیر انباشتهٔ خطاها در طول توالی حرکات مدل‌سازی شود. دوم آنکه، مسئله را در قالب رگرسیون متغیرهای پنهان مدل و از یک تابع هزینهٔ جریمهٔ یکنواختی[۴۹] برای اعمال محدودیت‌های فیزیکی استفاده می‌کند. با وجود این، مدل پیشنهادی در این مقاله، با تمرکز بر وابستگی‌های محلی مارکوفی، یک معماری پایه، منسجم و با پیچیدگی ساختاری و محاسباتی کمتر ارائه می‌دهد که به طور مؤثر چالش نبود برچسب برای زیرفعالیت‌ها را برطرف می‌کند. این روش توانسته است شکاف بین مدل‌های مستقل [۱۴] و مدل‌های جامع را به‌خوبی پر کند و با ایجاد توازنی میان دقت و هزینهٔ محاسباتی، به دقتی کاملاً رقابتی و قابل قبول دست یابد.

این نتایج بیانگر دقت زیاد مدل در پیش‌بینی رتبه‌بندی‌ها و امتیازهای کلی کیفیت فعالیت‌ها و نیز قابلیت تعمیم مناسب آن در مواجهه با نمونه‌های جدید است. این موفقیت نقش مؤثر بهره‌گیری از وابستگی زمانی میان زیرفعالیت‌ها را در ارتقای کیفیت ارزیابی عملکرد برجسته می‌کند. نتایج آزمایش‌ها در جدول (۱) نشان می‌دهد حتی با مدل‌سازی وابستگی محلی، دقت ارزیابی نسبت به روش‌های پایه [۱۴] بهبود یافته است.

جدول (۱): مقایسهٔ نتایج چارچوب پیشنهادی با روش‌های ارزیابی کیفیت فعالیت موجود. مقادیر پررنگ بهترین نتایج را نشان می‌دهند. توجه شود که اعداد مربوط به سلول‌های مشخص‌شده با علامت (–) در مقالهٔ اصلی گزارش نشده‌اند.

چارچوب	SRC	MSE	MED
[۱] Pose+DCT+SVR	۰/۵۳	–	–
Entropy feature ApEnFT [۲]	۰/۴۵	–	–
[۶] C3D+CNN	۰/۸۰	–	۷/۷۸
[۱۲] ScoringNet	۰/۸۴	–	۵/۳۶
[۹] S3D	۰/۸۶	۹۷/۴۶	۶/۹۰
[۴۰] Metric Learning	۰/۷۶	۱۰۵/۶۲	–
[۱۵] USDL	۰/۸۱	–	–
[۱۳] MRSM	۰/۸۸	۷۳/۹۲	–
[۵۲] PLSL	**۰/۹۱**	**۲۹/۴۲**	۴/۰۵
[۱۴] PSL	۰/۸۰	۸۶/۳۷	۷/۵۹
پیشنهادی (وابستگی محلی)	۰/۹۰	۳۲/۴۲	**۲/۸۳**



شکل (۳): نتایج امتیازدهی چارچوب پیشنهادی برای ۷۰ ویدئوی تست.

این مدل در مقایسه با رگرسیون تک‌مرحله‌ای و رویکردهای اخیر مبتنی بر بخش‌بندی ارزیابی شده است. برای بررسی دقیق‌تر عملکرد روش پیشنهادی، نمودار شکل (۳) امتیازهای پیش‌بینی‌شده را در مقایسه با امتیازهای واقعی برای ۷۰ نمونه ویدئوی تست نمایش می‌دهد. تطابق زیاد منحنی مدل پیشنهادی با مقادیر واقعی گواهی بر اثربخشی چارچوب طراحی‌شده در این پژوهش است. نتایج حاصل دقت مطلوب مدل را برای بیشتر نمونه‌های تست نشان می‌دهد و فقط در مواردی محدود انحراف قابل ملاحظه مشاهده می‌شود.

تحلیل کیفی نتایج نشان می‌دهد چارچوب پیشنهادی قادر است امتیازهایی واقع‌بینانه را در سطوح دشواری مختلف تولید کند. ارزیابی‌های مقایسه‌ای کارایی این مدل را در مواجهه با چالش‌هایی برجسته می‌کند که در رویکردهای مرتبط پیشین، مانند [۱۳]، به خطاهایی قابل ملاحظه منجر می‌شدند. برای نمونه، ویدئوی شمارهٔ ۲۳۵ از مجموعه‌تست که مربوط به اجرای یک «شیرجهٔ معکوس» با سطح دشواری نسبتاً زیاد ۳/٦ و امتیاز کلی واقعی ۲۱/٦ است، توسط رویکرد [۱۳] به طرزی قابل ملاحظه بیش از حد تخمین زده شده و مقدار ۷۳/۸٤ برای آن پیش‌بینی شده است؛ در حالی که چارچوب پیشنهادی، برای همین نمونه، امتیاز ٤٤/۹۰ را پیش‌بینی کرده که به مقدار واقعی نزدیک‌تر است (شکل (٤)). این‌گونه فعالیت‌ها (مانند توالی‌های شیرجهٔ معکوس یا بالانس بازو) به دلیل ساختار پیچیده و حساسیت زیاد، پیش‌بینی دقیق امتیاز کلی را دشوارتر می‌کنند و به‌شدت مستعد تخمین بیش از حد هستند.

بخشی از خطای مدل‌های پیشین به مسئلهٔ عدم توازن در توزیع برچسب‌های امتیاز کلی در مجموعه‌دادهٔ UNLV-Diving بازمی‌گردد. در این مجموعه، مقادیر امتیاز کلی از ۲۱/٦ تا ۱۰۲/٦ گسترده هستند، اما میانگین آنها تقریباً ۷۸ است. بیشتر نمونه‌ها دارای برچسب‌های امتیاز کلی بالا هستند، در حالی که نمونه‌هایی با امتیاز پایین به طرزی قابل ملاحظه کمتر در توزیع داده‌ها حضور دارند. این عدم تعادل، همراه با سطح دشواری بالای برخی از حرکات، موجب می‌شود رویکردهای قبلی در پیش‌بینی نمونه‌هایی با امتیاز پایین (مانند ویدئوی شمارهٔ ۲۳۵) عملکردی ضعیف‌تر داشته باشند. در مقابل، چارچوب پیشنهادی با بهره‌گیری از یادگیری متوالی شبه‌زیرامتیازها و درک وابستگی زمانی، از تخمین بیش از حد اجتناب کرده و توانسته است حتی برای نمونه‌های پیچیده با برچسب امتیاز پایین، پیش‌بینی‌ها را به‌خوبی با مقادیر واقعی هم‌راستا کند.

ویدئو شماره: ۲۳۵
سطح دشواری: ۳.۶
۷۳.۸۴ – ۴۴.۹۰ – ۲۱.۶

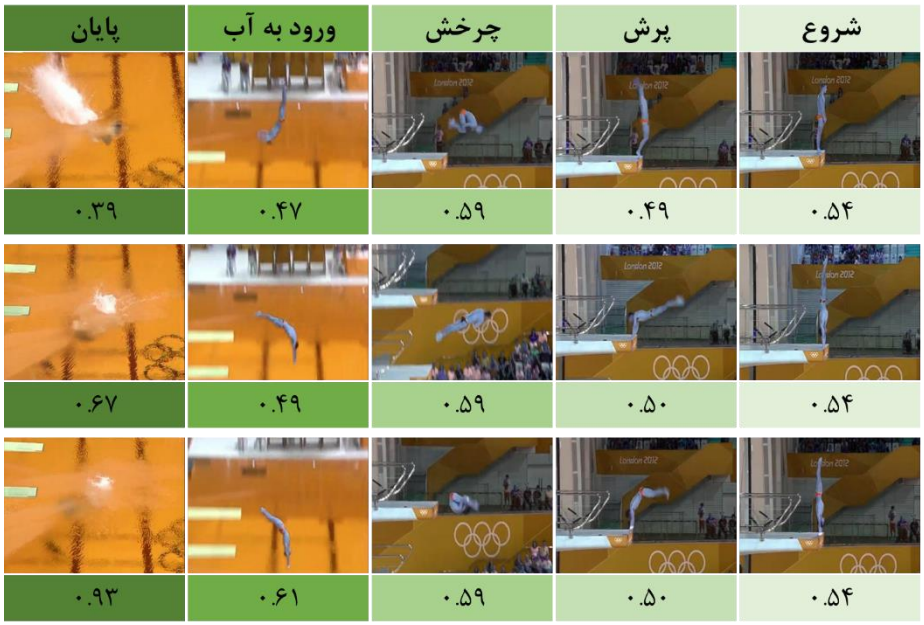
شکل (٤): نمونه‌ای با پیش‌بینی نادرست از شیرجهٔ معکوس با امتیاز کلی واقعی نسبتاً پایین نمایش داده شده است. مقدار سبز نمایانگر امتیاز کلی واقعی، مقدار قرمز نشان‌دهندهٔ امتیاز پیش‌بینی‌شده توسط رویکرد [۱۳] و مقدار آبی مربوط به امتیاز پیش‌بینی‌شده با استفاده از چارچوب پیشنهادی است.

۴-۳- بررسی عملکرد مدل در مراحل مختلف شیرجه

نتایج آزمایش‌ها نشان می‌دهد ادغام شبه‌زیرامتیازها دقت پیش‌بینی امتیاز کلی را بهبود می‌بخشد و امکان ارائهٔ بازخورد جزئی برای هر زیرفعالیت را فراهم می‌کند. تحلیل عملکرد چارچوب پیشنهادی در زیرفعالیت‌های مختلف ویدئوهای تست نشان می‌دهد زیرامتیازهای سه زیرفعالیت نخست معمولاً به یکدیگر نزدیک هستند. در مقابل، زیرفعالیت‌های ورود به آب و پایان بیشترین تنوع در امتیازها را در میان نمونه‌ها دارند. این موضوع بیانگر حساسیت زیاد مدل نسبت به این دو مرحله است؛ به گونه‌ای که پیش‌بینی‌ها در این زیرفعالیت‌ها بیشتر دچار نوسان می‌شوند. بنابراین، می‌توان آنها را به عنوان مراحل کلیدی در ارزیابی کیفیت اجرای شیرجه در نظر گرفت. این دو مرحله به ترتیب شامل ارزیابی موقعیت بدنی ورزشکار هنگام جمع‌کردن بالاتنه به سمت پاها، با تأکید بر کشیدگی کامل پاها و تماس بیشینهٔ بالاتنه با پاها و سنجش اندازهٔ پاشش آب هنگام ورود به آب هستند.

به منظور درک بهتر نقش مراحل مختلف اجرای شیرجه در شکل‌گیری امتیاز نهایی، شکل (۵) به تحلیل پنج زیرفعالیت شیرجه در سه نمونهٔ منتخب اختصاص یافته است. این سه نمونه به ترتیب نمایندهٔ اجراهایی با امتیاز کلی بالا، متوسط و پایین هستند. مقادیر واقعی امتیاز کلی آنها به ترتیب ۹۹/۷۵، ۷۷/۷ و ۴۲/۹ هستند و مقادیر پیش‌بینی‌شده توسط چارچوب پیشنهادی نیز به ترتیب ۹۸/۳۱، ۷۶/۹۳ و ۳۸/۸۵ به دست آمده‌اند. دو نمونه از این مجموعه (نمونه‌های ۱۰۲ و ۲۳۸) دارای نوعی خاص از شیرجه با سطح دشواری نسبتاً بالا هستند که در آن، ورزشکار ابتدا روی دستان خود روی سکوی شیرجه می‌ایستد و سپس شیرجه را انجام می‌دهد. در شکل (۵)، مقادیر سبز نمایانگر امتیازهای واقعی و مقادیر آبی نشان‌دهندهٔ امتیازهای پیش‌بینی‌شده توسط چارچوب پیشنهادی هستند. برای هر یک از زیرفعالیت‌ها، یک فریم نماینده و مقادیر زیرامتیاز پیش‌بینی‌شده برای این سه نمونه در تصویر نمایش داده شده‌اند. این نمایش بصری امکان مقایسه‌ای دقیق‌تر از کیفیت پیش‌بینی در مراحل مختلف شیرجه را فراهم می‌کند. به‌وضوح، در نمونهٔ شمارهٔ ۲۳۸ که امتیاز کلی بالاتری دارد، مرحلهٔ چهارم با زیرامتیاز ۰/۶۱ بالاترین ارزیابی را دریافت کرده است. در این مرحله، ورزشکار از وضعیت بدنی استانداردی برخوردار است و زاویهٔ میان بالاتنه و پاها کمتر است؛ این مسئله با درک عمومی و همچنین دانش تخصصی داوری در شیرجه نیز هم‌خوانی دارد. در مرحلهٔ پنجم که تمرکز بر لندازهٔ پاشش آب به جای تحلیل حرکات بدن است، این نمونه زیرامتیاز نسبتاً بالایی برابر ۰/۹۳ به دست آورده است. این مقدار نشان‌دهندهٔ کوچک‌ترین میزان پاشش آب و در نتیجه، اجرای دقیق‌تر و فرود تمیزتر در آب است. در مقابل، نمونهٔ شمارهٔ ۳۲۴ که امتیاز کلی پایین‌تری دریافت کرده است، در تمامی مراحل، زیرامتیازهای پایین‌تر کسب کرده است. به ویژه در مرحلهٔ ورود به آب، زیرامتیاز ۰/۴۷ و در مرحلهٔ پایان ۰/۳۹ ثبت شده که نشان‌دهندهٔ کیفیت پایین‌تر فرود و در نتیجه پاشش زیاد آب است.

این تحلیل مؤید آن است که چارچوب پیشنهادی نه فقط توانایی پیش‌بینی امتیاز کلی را دارد، بلکه قادر به ارزیابی دقیق عملکرد ورزشکار در مراحل جزئی نیز هست. چنین قابلیتی به مربیان و داوران این امکان را می‌دهد تا علت اصلی افت کیفیت در هر اجرای خاص را به‌راحتی شناسایی و برطرف کنند. همچنین ورزشکاران را قادر می‌کند تا بر بهبود جنبه‌های خاص عملکرد خود تمرکز کنند و در نتیجه، مهارت‌های فنی و نتایج رقابتی خود را ارتقا دهند.



شکل (۵): سه نمونه با امتیازهای متنوع. مقدار سبز نشان‌دهندهٔ برچسب امتیاز کلی و مقدار آبی مربوط به امتیاز پیش‌بینی‌شده با استفاده از چارچوب پیشنهادی است.

۳-۵- تحلیل تفسیرپذیری ارزیابی کیفیت در سطح زیرفعالیت‌ها و ارزیابی معنایی شبه‌زیرامتیازها در مجموعه‌دادهٔ UNLV-Diving

یکی از چالش‌های اصلی در رویکردهای مبتنی بر یادگیری با نظارت ضعیف، اثبات معناداربودن متغیرهای پنهان یا همان شبه‌زیرامتیازهای تولیدشده است. برای بررسی اینکه آیا رویکرد پیشنهادی «وابستگی متوالی» توانسته است منطق داوری انسانی را به‌درستی مدل‌سازی کند، یک تحلیل توزیع طبقه‌بندی‌شده[۵۰] بر روی نمونه‌های آزمون در مجموعه‌دادهٔ UNLV-Diving انجام شد. به منظور فراهم‌سازی امکان مقایسهٔ بصری و تحلیلی، ویدئوهای مجموعه‌آزمون بر اساس برچسب امتیاز کلی واقعی به سه دستهٔ کیفیتی تقسیم شدند: کیفیت ضعیف (امتیاز کمتر از ۷۰)، کیفیت متوسط (امتیاز بین ۷۰ تا ۸۵) و کیفیت عالی (امتیاز بیشتر از ۸۵). سپس توزیع شبه‌زیرامتیازهای پیش‌بینی‌شده برای هر ۵ زیرفعالیت ورزش شیرجه تحلیل شد. این دسته‌بندی امکان مقایسهٔ واضحِ عملکردِ شبکه در زیرفعالیت‌های مختلف (شامل ۱: شروع، ۲: پرش، ۳: چرخش، ٤: ورود به آب و ۵: پایان) را برای سطوح مختلفِ مهارت ورزشکار فراهم می‌کند.

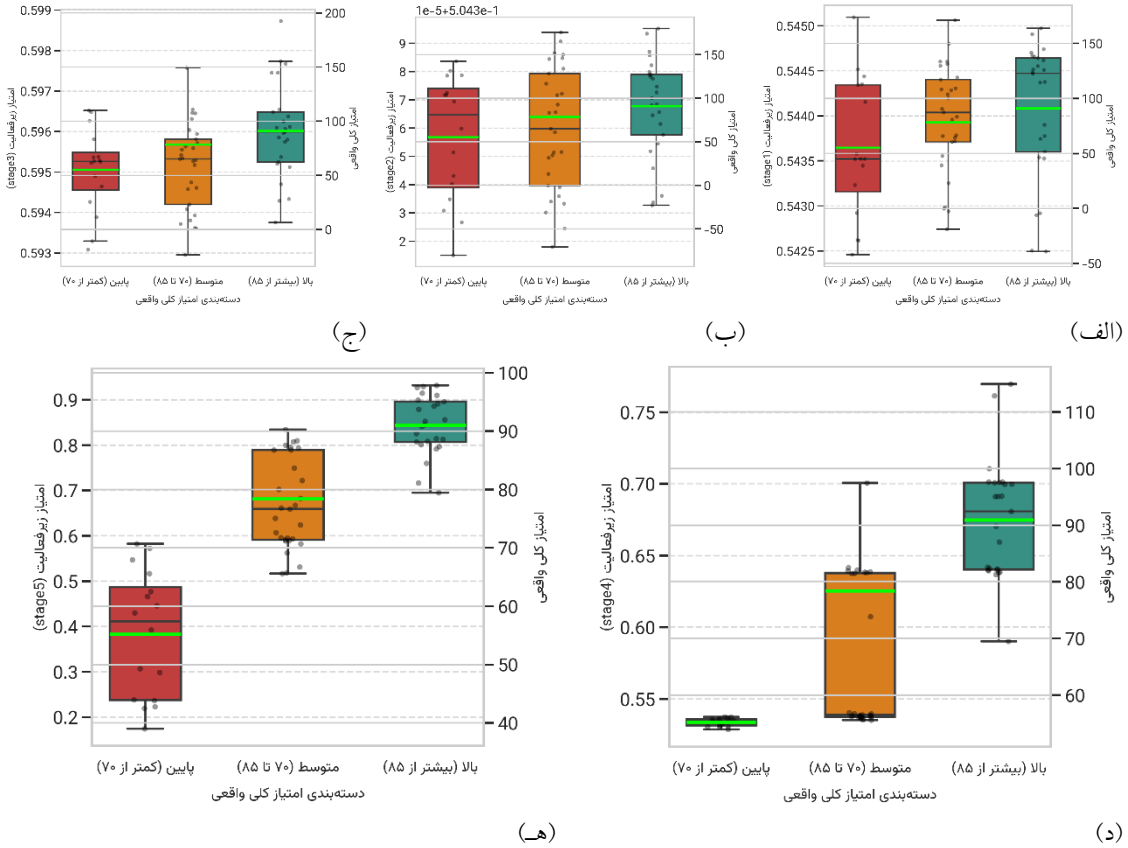
در راستای ارائهٔ درک بصری دقیق‌تر، نمودارهای جعبه‌ای[۵۱] در شکل (٦) با استفاده از طراحی دومحوره[۵۲] رسم شده‌اند. محور عمودی سمت چپ نشان‌دهندهٔ توزیع «شبه‌زیرامتیازهای پیش‌بینی‌شده» است، در حالی که محور عمودی سمت راست و خط ممتد سبزرنگ نمایانگر «میانگین امتیاز کلی واقعی» در هر یک از دسته‌های کیفی است. همان‌طور که در نمودارها مشاهده می‌شود، نتایج حاکی از وجود یک تناسب کاملاً منطقی و مستقیم میان شبه‌زیرامتیازهای پیش‌بینی‌شده و کیفیت واقعی عملکرد ورزشکار در تمامی مراحل است. به عبارت دیگر، با افزایش سطح مهارت و امتیاز کلی، میانهٔ شبه‌زیرامتیازها در هر زیرفعالیت نیز به طور پیوسته افزایش می‌یابد. این هم‌راستاییِ دقیق به صورت کمّی و بصری ثابت می‌کند شبه‌زیرامتیازهای یادگرفته‌شده بازنمایی‌کننده‌ای بسیار خوب از امتیاز کلی هستند و شبکه توانسته است سهم هر حرکت را در نمرهٔ نهایی به‌درستی تخمین بزند.

با بررسی روند هر ۵ زیرفعالیت، به صورت بصری کاملاً مشهود است که دو زیرفعالیت نهایی، یعنی مرحلهٔ ٤ (ورود به آب) و مرحلهٔ ۵ (پایان)، نقشی بسیار بارزتر و تأثیرگذارتر در محاسبهٔ امتیاز کلی ایفا می‌کنند. برای نمونه، در زیرفعالیت پنجم (پایان)، تفکیکی بسیار واضح میان کلاس‌ها دیده

می‌شـــود: گروه بـا کیفیـت پـایین حول میـانهٔ ۰/٤۱ متمرکز شـــده‌اند، در حالی که گروه با کیفیت بالا به طرزی قابل ملاحظه میانهٔ بالاتری در حدود ۰/۸۵ را ثبت کرده‌اند.

این رفتار مدل در تطابق کامل با قوانین فنی و داوری فدراسـیون جهانی شـنا (FINA) قرار دارد. تحلیل‌ها نشـان می‌دهند شـبه‌زیرامتیازهای مربوط به مراحل «ورود به آب» و «پایان» بیشـترین قدرت تفکیک را میان سـطوح کیفی از خود نشان می‌دهند. شیرجه‌های با کیفیت کم در این مراحل دارای پراکندگی (واریانس) بالا و میانهٔ بسیار پایینی هستند؛ رفتاری که دقیقاً بازتاب‌دهندهٔ جریمه‌های سـنگینِ داوران به دلیل خطای پاشـش شـدید آب اسـت. در مقابل، شـیرجه‌های با کیفیت زیاد دارای توزیعی متراکم، فشـرده و با مقادیر بالا هسـتند که نشـان‌دهندهٔ اجرای مداوم و تکنیک ورود تمیز و بدون پاشـش آب اسـت که همواره امتیازات طلایی داوران متخصص را به همراه دارد.

این تحلیل کمّی و بصـری ثابت می‌کند چارچوب پیشنهادی ما صرفاً یک جعبه‌سیاه نبوده و فقط به بهینه‌سازی ریاضـی بسـنده نکرده اسـت، بلکه با بهره‌گیری از مکانیسـم ادغام شـبه‌زیرامتیازِ مرحلهٔ قبل با ویژگی‌های مرحلهٔ فعلی، موفق شده است وزن، اهمیت و زنجیرهٔ علّی‌ومعلولی مراحل مختلف اجرای حرکت را دقیقاً هم‌راستا با معیارهای ارزیابی داوران متخصص انسانی یاد بگیرد.



شکل (٦): تحلیل توزیع طبقه‌بندی‌شدهٔ شبه‌زیرامتیازهای تولیدشده در سه دستهٔ کیفی مختلف (پایین: کمتر از ۷۰، متوسط: بین ۷۰ تا ۸۵، و بالا: بیشتر از ۸۵) بر روی مجموعه‌دادهٔ UNLV-Diving. روند صعودی و یکنواخت میانهٔ امتیازها از دستهٔ با کیفیت کم به دستهٔ با کیفیت زیاد در تمامی زیرفعالیت‌ها به صورت کمّی انسجام معنایی شبه‌زیرامتیازهای یادگرفته‌شده را اثبات می‌کند. گفتنی است، زیرفعالیت‌های ورود به آب (د) و پایان (هـ) بیشترین پراکندگی (واریانس) را در گروه با کیفیت کم نشان می‌دهند که به‌خوبی بازتاب‌دهندهٔ جریمه‌های سنگین مرتبط با خطای پاشش آب است؛ در حالی که شیرجه‌های باکیفیت توزیعی متراکم با مقادیر بالا را نشان می‌دهند که گواهی بر اجرای پایدار و بی‌نقص حرکت است.

۴- جمع‌بندی و نتیجه‌گیری

در این پژوهش، یک چارچوب نوین برای ارزیابی کیفیت فعالیت ارائه شد که با بهره‌گیری از ساختار زیرفعالیت‌ها و یادگیری متوالی شبه‌زیرامتیازها، به صورت مؤثر به چالش نبود برچسب‌های صریح در سطح زیرفعالیت پاسخ می‌دهد. مدل پیشنهادی با ترکیب بخش‌بندی معنایی زمانی، استخراج ویژگی‌های مکانی-زمانی و تقویت گام‌به‌گام نمایش زیرفعالیت‌ها، توانسته است وابستگی‌های زمینه‌ای درون‌مرحله‌ای و روابط زمانی میان زیرفعالیت‌ها را به‌دقت مدل‌سازی کند. با استفاده از ویژگی‌های مکانی-زمانی و برچسب امتیاز کلی، شبه‌زیرامتیازهای اولیه برای زیرفعالیت‌ها در داده‌های آموزشی تولید می‌شوند. سپس، چارچوب پیشنهادی با بهره‌گیری هم‌زمان از امتیاز کلی و شبه‌زیرامتیاز مرحلهٔ قبل، بازنمایی دقیق‌تری از زیرفعالیت‌ها ایجاد می‌کند. این طراحی به مدل اجازه می‌دهد تا علاوه بر پیش‌بینی دقیق امتیاز کلی، بازخوردهای هدفمند و قابل تفسیر برای هر مرحله از اجرای فعالیت ارائه دهد؛ به طوری که تحلیل‌های بصری و آماری اثبات می‌کنند این زیرامتیازها انطباق بسیار زیادی با معیارهای داوری انسانی (مانند قوانین داوری FINA) دارند.

نتایج مقایسه‌ای نشان می‌دهد چارچوب پیشنهادی، به ویژه در سناریوهای ورزشی که در آنها زیرفعالیت‌ها به طور زنجیره‌وار بر کیفیت نهایی تأثیر می‌گذارند، عملکردی مطلوب، رقابتی و بهبودیافته نسبت به روش‌های پایه (که زیرفعالیت‌ها را مستقل فرض می‌کنند) ارائه می‌دهد. بهبود معیارهایی مانند SRC، MED و MSE، اثربخشی این معماری را تأیید می‌کند. در ادامه، توسعهٔ این چارچوب برای کاربرد در حوزه‌های ورزشی متنوع و بهره‌گیری از مجموعه‌داده‌های غنی‌تر در دستور کار قرار دارد. تمرکز اصلی، ایجاد مجموعه‌داده‌هایی با برچسب‌های کیفیت در سطح زیرفعالیت است تا امکان اعتبارسنجی مستقیم زیرامتیازهای پیش‌بینی‌شده فراهم شود. همچنین، بهبود روش‌های یادگیری بازنمایی ویدئویی نیز می‌تواند استحکام چارچوب را در شرایط پیچیدهٔ بصری افزایش دهد.

مراجع

[1] H. Pirsiavash, C. Vondrick, A. Torralba, “Assessing the Quality of Actions”, in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Vol. 8694 LNCS, No. PART 6, pp. 556–571, 2014. https://doi.org/10.1007/978-3-319-10599-4_36

[2] V. Venkataraman, I. Vlachos, P. Turaga, “Dynamical Regularity for Action Analysis”, pp. 67.1-67.12, 2015. https://doi.org/10.5244/c.29.67

[3] H. Li, Q. Lei, H. Zhang, J. Du, S. Gao, “Skeleton-based deep pose feature learning for action quality assessment on figure skating videos”, J. Vis. Commun. Image Represent., Vol. 89, p. 103625, Nov. 2022. https://doi.org/10.1016/j.jvcir.2022.103625

[4] P. Parmar, B. T. Morris, “Learning to Score Olympic Events”, in 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), IEEE, pp. 76–84, Jul. 2017. https://doi.org/10.1109/CVPRW.2017.16

[5] D. Tran, L. Bourdev, R. Fergus, L. Torresani, M. Paluri, “Learning Spatiotemporal Features with 3D Convolutional Networks”, in 2015 IEEE International Conference on Computer Vision (ICCV), IEEE, pp. 4489–4497, Dec. 2015. https://doi.org/10.1109/ICCV.2015.510

[6] Y. Li, X. Chai, X. Chen, “End-To-End Learning for Action Quality Assessment”, in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Vol. 11165 LNCS, pp. 125–134, 2018. https://doi.org/10.1007/978-3-030-00767-6_12

[7] P. Parmar, B. T. Morris, “Action quality assessment across multiple actions”, in Proceedings - 2019 IEEE Winter Conference on Applications of Computer Vision, WACV 2019, pp. 1468–1476, 2019. https://doi.org/10.1109/WACV.2019.00161

[8] P. Parmar, B. T. Morris, “What and how well you performed? a multitask learning approach to action quality assessment”, in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 304–313, 2019. https://doi.org/10.1109/CVPR.2019.00039

[9] X. Xiang, Y. Tian, A. Reiter, G. D. Hager, T. D. Tran, “S3D: Stacking Segmental P3D for Action Quality Assessment”, in Proceedings - International Conference on Image Processing, ICIP, pp. 928–932, 2018. https://doi.org/10.1109/ICIP.2018.8451364

[10] C. Lea, M. D. Flynn, R. Vidal, A. Reiter, G. D. Hager, “Temporal Convolutional Networks for Action Segmentation and Detection”, in 2017 IEEE Conference on Computer Vision and Pattern

Assist. Radiol. Surg., Vol. 16, No. 9, pp. 1595–1605, 2021. https://doi.org/10.1007/s11548-021-02448-4

[21] S. Farabi, H. Himel, F. Gazzali, M. B. Hasan, M. H. Kabir, M. Farazi, "Improving Action Quality Assessment Using Weighted Aggregation", in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), pp. 576–587, 2022. https://doi.org/10.1007/978-3-031-04881-4_46

[22] Y. Zhang, W. Xiong, S. Mi, "Learning time-aware features for action quality assessment", Pattern Recognit. Lett., Vol. 158, pp. 104–110, Jun. 2022. https://doi.org/10.1016/j.patrec.2022.04.015

[23] S. Zahan, G. M. Hassan, A. Mian, "Learning Sparse Temporal Video Mapping for Action Quality Assessment in Floor Gymnastics", arXiv Prepr. arXiv2301.06103, 2023. [Online]. Available: http://arxiv.org/abs/2301.06103

[24] A. Iyer, M. Alali, H. Bodala, S. Vaidya, "Action Quality Assessment using Transformers", arXiv Prepr. arXiv2207.12318, Jul. 2022. [Online]. Available: http://arxiv.org/abs/2207.12318

[25] A. Xu, L. A. Zeng, W. S. Zheng, "Likert Scoring with Grade Decoupling for Long-term Action Assessment", in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 3222–3231, 2022. https://doi.org/10.1109/CVPR52688.2022.00323

[26] S. Wang, Di. Yang, P. Zhai, C. Chen, L. Zhang, "TSA-Net: Tube Self-Attention Network for Action Quality Assessment", in MM 2021 - Proceedings of the 29th ACM International Conference on Multimedia, pp. 4902–4910, 2021. https://doi.org/10.1145/3474085.3475438

[27] M.-Z. Li, H.-B. Zhang, L.-J. Dong, Q. Lei, J.-X. Du, "Gaussian guided frame sequence encoder network for action quality assessment", Complex Intell. Syst., Vol. 9, No. 2, pp. 1963–1974, Apr. 2023. https://doi.org/10.1007/s40747-022-00892-6

[28] K. Zhou, Y. Ma, H. P. H. Shum, X. Liang, "Hierarchical Graph Convolutional Networks for Action Quality Assessment", IEEE Trans. Circuits Syst. Video Technol., pp. 1–1, 2023. https://doi.org/10.1109/TCSVT.2023.3281413

[29] C. Han, F. Shen, L. Chen, X. Lian, H. Gou, H. Gao, "MLA-LSTM: A Local and Global Location Attention LSTM Learning Model for Scoring Figure Skating", Systems, Vol. 11, No. 1, p. 21, Jan. 2023. https://doi.org/10.3390/systems11010021

[30] B. Zhang, J. Chen, Y. Xu, H. Zhang, X. Yang, X. Geng, "Auto-encoding score distribution regression for action quality assessment", Neural Recognition (CVPR), IEEE, pp. 1003–1012, Jul. 2017. https://doi.org/10.1109/CVPR.2017.113

[11] P. Parmar, B. Morris, "HalluciNet-ing Spatiotemporal Representations Using a 2D-CNN", Signals, Vol. 2, No. 3, pp. 604–618, 2021. https://doi.org/10.3390/signals2030037

[12] Y. Li, X. Chai, X. Chen, "ScoringNet: Learning Key Fragment for Action Quality Assessment with Ranking Loss in Skilled Sports", in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), pp. 149–164, 2019. https://doi.org/10.1007/978-3-030-20876-9_10

[13] L. J. Dong, H. B. Zhang, Q. Shi, Q. Lei, J. X. Du, S. Gao, "Learning and fusing multiple hidden substages for action quality assessment", Knowledge-Based Syst., Vol. 229, 2021. https://doi.org/10.1016/j.knosys.2021.107388

[14] H.-B. Zhang, L.-J. Dong, Q. Lei, L.-J. Yang, J.-X. Du, "Label-reconstruction-based pseudo-subscore learning for action quality assessment in sporting events", Appl. Intell., Vol. 53, No. 9, pp. 10053–10067, May 2023. https://doi.org/10.1007/s10489-022-03984-5

[15] Y. Tang, Z. Ni, J. Zhou, D. Zhang, J. Lu, Y. Wu, et al., "Uncertainty-Aware Score Distribution Learning for Action Quality Assessment", in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 9836–9845, 2020. https://doi.org/10.1109/CVPR42600.2020.00986

[16] C. Zhou, Y. Huang, "Uncertainty-Driven Action Quality Assessment", arXiv Prepr. arXiv2207.14513, Vol. 14, No. 8, pp. 1–10, Jul. 2022. [Online] Available: http://arxiv.org/abs/2207.14513

[17] C. Xu, Y. Fu, B. Zhang, Z. Chen, Y. G. Jiang, X. Xue, "Learning to Score Figure Skating Sport Videos", IEEE Trans. Circuits Syst. Video Technol., Vol. 30, No. 12, pp. 4578–4590, 2020. https://doi.org/10.1109/TCSVT.2019.2927118

[18] L. A. Zeng, F. T. Hong, W. S. Zheng, Q. Z. Yu, W. Zeng, Y. W. Wang, et al., "Hybrid Dynamic-static Context-aware Attention Network for Action Assessment in Long Videos", in MM 2020 - Proceedings of the 28th ACM International Conference on Multimedia, pp. 2526–2534, 2020. https://doi.org/10.1145/3394171.3413560

[19] Q. Lei, H. Zhang, J. Du, "Temporal attention learning for action quality assessment in sports video", Signal, Image Video Process., Vol. 15, No. 7, pp. 1575–1583, 2021. https://doi.org/10.1007/s11760-021-01890-w

[20] T. Wang, M. Jin, M. Li, "Towards accurate and interpretable surgical skill assessment: a video-based method for skill score prediction and guiding feedback generation", Int. J. Comput.

[40] H. Jain, G. Harit, A. Sharma, "Action Quality Assessment Using Siamese Network-Based Deep Metric Learning", IEEE Trans. Circuits Syst. Video Technol., Vol. 31, No. 6, pp. 2260–2273, 2021. https://doi.org/10.1109/TCSVT.2020.3017727

[41] D. Yi, Z. Lei, S. Liao, S. Z. Li, "Deep metric learning for person re-identification", in Proceedings - International Conference on Pattern Recognition, pp. 34–39, 2014. https://doi.org/10.1109/ICPR.2014.16

[42] X. Yu, Y. Rao, W. Zhao, J. Lu, J. Zhou, "Group-aware Contrastive Regression for Action Quality Assessment", in Proceedings of the IEEE International Conference on Computer Vision, pp. 7899–7908, 2021. https://doi.org/10.1109/ICCV48922.2021.00782

[43] J. Xu, Y. Rao, X. Yu, G. Chen, J. Zhou, J. Lu, "FineDiving: A Fine-grained Dataset for Procedure-aware Action Quality Assessment", in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 2939–2948, 2022. https://doi.org/10.1109/CVPR52688.2022.00296

[44] Y. Bai, D. Zhou, S. Zhang, J. Wang, E. Ding, Y. Guan, et al., "Action Quality Assessment with Temporal Parsing Transformer", in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Vol. 13664 LNCS, pp. 422–438, 2022. https://doi.org/10.1007/978-3-031-19772-7_25

[45] M. Li, H. B. Zhang, Q. Lei, Z. Fan, J. Liu, J. X. Du, "Pairwise Contrastive Learning Network for Action Quality Assessment", in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), pp. 457–473, 2022. https://doi.org/10.1007/978-3-031-19772-7_27

[46] L. Liu, P. Zhai, D. Zheng, Y. Fang, "Multi-Stage Action Quality Assessment Method", in 2023 4th International Conference on Control, Robotics and Intelligent System, New York, NY, USA: ACM, pp. 116–122, Aug. 2023. https://doi.org/10.1145/3622896.3622916

[47] A. Zia, I. Essa, "Automated surgical skill assessment in RMIS training", Int. J. Comput. Assist. Radiol. Surg., Vol. 13, No. 5, pp. 731–739, May 2018, https://doi.org/10.1007/s11548-018-1735-5

[48] A. Zia, Y. Sharma, V. Bettadapura, E. L. Sarin, I. Essa, "Video and accelerometer-based motion analysis for automated surgical skills assessment", Int. J. Comput. Assist. Radiol. Surg., Vol. 13, No. 3, pp. 443–455, Mar. 2018. https://doi.org/10.1007/s11548-018-1704-z

[49] J. Carreira, A. Zisserman, "Quo Vadis, action recognition? A new model and the kinetics Comput. Appl., Oct. 2023. https://doi.org/10.1007/s00521-023-09068-w

[31] S. Zhang, W. Dai, S. Wang, X. Shen, J. Lu, J. Zhou, et al., "LOGO: A Long-Form Video Dataset for Group Action Quality Assessment", in 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, pp. 2405–2414, Jun. 2023. https://doi.org/10.1109/CVPR52729.2023.00238

[32] H. Matsuyama, N. Kawaguchi, B. Y. Lim, "IRIS: Interpretable Rubric-Informed Segmentation for Action Quality Assessment", in Proceedings of the 28th International Conference on Intelligent User Interfaces, New York, NY, USA: ACM, pp. 368–378, Mar. 2023. https://doi.org/10.1145/3581641.3584048

[33] G. I. Parisi, S. Magg, S. Wermter, "Human motion assessment in real time using recurrent self-organization", in 2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN), IEEE, pp. 71–76, Aug. 2016. https://doi.org/10.1109/ROMAN.2016.7745093

[34] S. J. Zhang, J. H. Pan, J. Gao, W. S. Zheng, "Semi-Supervised Action Quality Assessment with Self-Supervised Segment Feature Recovery", IEEE Trans. Circuits Syst. Video Technol., 2022. https://doi.org/10.1109/TCSVT.2022.3143549

[35] L.-A. Zeng, W.-S. Zheng, "Multimodal Action Quality Assessment", IEEE Trans. Image Process., Vol. 33, pp. 1600–1613, 2024. https://doi.org/10.1109/TIP.2024.3362135

[36] T. Nagai, S. Takeda, S. Suzuki, H. Seshimo, "MMW-AQA: Multimodal In-the-Wild Dataset for Action Quality Assessment", IEEE Access, Vol. 12, pp. 92062–92072, 2024. https://doi.org/10.1109/ACCESS.2024.3423462

[37] H. Doughty, D. Damen, W. Mayol-Cuevas, "Who's Better? Who's Best? Pairwise Deep Ranking for Skill Determination", in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE, pp. 6057–6066, Jun. 2018. https://doi.org/10.1109/CVPR.2018.00634

[38] H. Doughty, W. Mayol-Cuevas, Di. Damen, "The Pros and Cons: Rank-Aware Temporal Attention for Skill Determination in Long Videos", in 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, pp. 7854–7863, Jun. 2019. https://doi.org/10.1109/CVPR.2019.00805

[39] Z. Li, Y. Huang, M. Cai, Y. Sato, "Manipulation-skill assessment from videos with spatial attention network", in Proceedings - 2019 International Conference on Computer Vision Workshop, ICCVW 2019, IEEE, pp. 4385–4395, Oct. 2019. https://doi.org/10.1109/ICCVW.2019.00539

[50] Z. Qiu, T. Yao, T. Mei, “Learning Spatio-Temporal Representation with Pseudo-3D Residual Networks”, in Proceedings of the IEEE International Conference on Computer Vision, pp. 5534–5542, 2017. https://doi.org/10.1109/ICCV.2017.590

[51] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, et al., “Automatic differentiation in pytorch”, 2017. Available: https://openreview.net/forum?id=BJJsrmfCZ

[52] M. Mazruei, E. Fazl-Ersi, A. Vahedian, A. Harati, “Context-aware action quality assessment using latent regression-based progressive sub-action learning”, Comput. Knowl. Eng., accepted manuscript, Feb. 17, 2026. https://doi.org/10.22067/cke.2026.96007.1177

---

[1] Action Quality Assessment (AQA)
[2] objectively
[3] Human Action Recognition (HAR)
[4] Regression-based
[5] Pairwise-comparison based
[6] Grading-based
[7] ground-truth scores
[8] Support Vector Regression (SVR)
[9] Discrete Cosine Transformation (DCT)
[10] approximate entropy
[11] splash
[12] ranking loss
[13] Encoder-Decoder TCN (ED-TCN)
[14] sub-actions
[15] Conditional Variational Auto-Encoders
[16] Temporal Attention
[17] self-attentive LSTM
[18] multi-scale convolutional skip LSTM
[19] Graph Convolutional Networks (GCN)
[20] pooling
[21] reinforcement learning
[22] discriminative non-local attention
[23] cross-attention
[24] pairwise deep ranking
[25] Siamese two-stream CNN
[26] spatial attention
[27] attention pooling
[28] temporal aggregation
[29] Group-Aware Regression Tree
[30] sub-stages
[31] step transitions
[32] down-up
[33] linear
[34] max pooling
[35] binary cross-entropy loss
[36] take-off
[37] rotation
[38] somersault
[39] experimental setup
[40] dropping
[41] regularization

[42] weight decay
[43] drop-out
[44] average pooling
[45] adaptive moment estimation (Adam)
[46] learning rate
[47] Local Context
[48] Long-range Cumulative Context
[49] Monotonic Penalty Loss
[50] Stratified Distribution Analysis
[51] Boxplots
[52] Dual-axis