مجله کنترل
ISSN (print) 2008-8345
ISSN (online) 2538-3752
مقاله پژوهشی
جلد۱۷، شماره ۲، تابستان ۱۴۰۲، صفحه ۴۶-۲۵

پردازش تصویر و بینایی ماشین در جراحی و آموزش آن

محمدجواد احمدی[۱]، محمدسینا اله‌کرم[۲]، پریسا عبدی[۳]، سیدفرزاد محمدی[۴]، حمیدرضا تقی راد[۵]

[۱] دانشجو دکتری مهندسی برق، گروه کنترل، دانشگاه صنعتی خواجه نصیرالدین طوسی،تهران، ایران mjahmadi@email.kntu.ac.ir

[۲] کارشناسی ارشد مهندسی برق، گروه کنترل، دانشگاه صنعتی خواجه نصیرالدین طوسی،تهران، ایران msina_alahkaram@email.kntu.ac.ir

[۳] استادیار، مرکز تحقیقات چشم‌پزشکی ترجمانی، بیمارستان فارابی، دانشگاه علوم پزشکی تهران ، ایران pabdi@tums.ac.ir

[۴] استاد، مرکز تحقیقات چشم‌پزشکی ترجمانی، بیمارستان فارابی، دانشگاه علوم پزشکی تهران، ایران sfmohammadi@tums.ac.ir

[۵] استاد، دانشکده مهندسی برق، گروه کنترل، دانشگاه صنعتی خواجه نصیرالدین طوسی ،تهران، ایران taghirad@kntu.ac.ir

دریافت: ۱۴۰۲/۰۵/۱۰ پذیرش: ۱۴۰۲/۰۶/۲۵

چکیده: با پیشرفت‌های هوش مصنوعی در در دههٔ اخیر، استفاده از دادهٔ تصویری و ویدیویی و فن‌آوری‌های مبتنی بر پردازش تصویر برای خودکارسازی روش‌های جراحی و آموزش آن، رونق یافته است. امروزه در بیش‌ترِ اتاق‌های عمل از یک یا چند دوربین و یا دستگاه ثبت اطلاعات استفاده می‌شود تا دادهٔ مهم پزشکی برای انجام تحلیل‌های بعدی ذخیره شوند. از این اطلاعات تصویری می‌توان برای طراحی و توسعهٔ سامانه‌های خودکار هدایت تصویری با هدف کمک به پزشک متخصص حین جراحی و آموزش آن استفاده کرد. هم‌چنین، این سامانه‌ها می‌توانند به‌عنوان مغز ابزارهای رباتیکی کمک‌جراح فعالیت کنند. یک سامانهٔ هدایت تصویری جراحی نیاز به قسمت‌های مختلفی دارد. از مهم‌ترین این قسمت‌ها می‌توان به تشخیص، بخش‌بندی و ردیابی ابزارها و نواحی مهم جراحی، تشخیص مراحل، حرکات و ژست‌ها، و تشخیص مهارت‌های جراحی اشاره کرد. خودکارکردن این بخش‌ها با استفاده از پردازش تصویر و بینایی ماشین کمک می‌کند، تا سامانه درک مستقل و عمیقی از صحنهٔ جراحی داشته باشد. در این مقاله ابتدا تعدادی از مجموعه‌داده‌های تصویری مهم مربوط به جراحی معرفی شده، و سپس شماری از پژوهش‌های اثرگذار در زمینهٔ پردازش تصویر و بینایی ماشین در کاربردهای ذکرشده با هدف ایجاد اجزای یک سامانهٔ خودکار هدایت تصویری جراحی، معرفی شده و زمینه های تحقیقاتی پیش رو معرفی می‌شوند.

کلمات کلیدی: بینایی ماشین در جراحی، سامانهٔ هدایت تصویری جراحی، تشخیص و بخش‌بندی، تشخیص مرحله، ارزیابی مهارت.

Image Processing and Machine Vision in Surgery and Its Training

Mohammad Javad Ahmadi, Mohammad Sina Allahkaram, Parisa Abdi, Seyed-Farzad Mohammadi, Hamid D. Taghirad

Abstract: Due to recent advancements in artificial intelligence, the utilization of image and video data along with image processing technologies has been significantly used in automating surgical procedures and their training. Today, one or more cameras or imaging devices are commonly used in most operating rooms to capture crucial medical information for further analysis. These visual data can be leveraged to design and develop automated image-guided systems aimed at assisting specialized surgeons during procedures and facilitating their training. Furthermore, these systems can serve as the cognitive knowledge for assisting surgical robots. An image-guided surgical system comprises of various components, among which the identification, segmentation, and tracking of surgical instruments and critical anatomical regions, phase recognition, motion analysis, and gesture recognition, and more importantly as surgical skill assessment, are the most crucial components. Automating these components by the use of image processing and machine vision will certainly contribute to the system's deep and independent understanding of the surgical scene. This article introduces the important datasets relevant to image processing and machine vision in surgical applications. Subsequently, several research endeavors in the realm of image processing, machine

vision, on the above-mentioned applications are presented, and some prospect research areas are introduced.

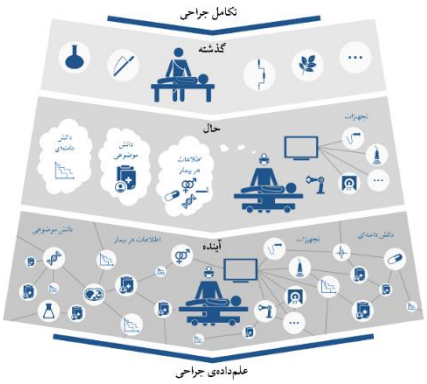
Keywords: Machine Vision in Surgery, Image-guided surgical system, Detection and Segmentation, Phase Recognition, Skill Assessment

۱- مقدمه

جراحی، حرفه‌ای است که از طریق تهاجم بدنی، در درمان بیماران به کار گرفته می شود [۱]. روش‌های جراحی نیاز به مهارت بالا، کنترل هماهنگی حرکات چشم و دست دارد. هم‌چنین نظر به خطرات جانی احتمالی برای بیماران بسیار مهم است که جراحان کارآموز در طول انجام عمل جراحی آموزشی بر روی بیماران، به درستی آموزش ببینند. شیوهٔ انجام و آموزش جراحی همانند بسیاری از زمینه‌های دیگر، به‌طور قابل توجهی در طول سالهای گذشته تکامل یافته است. در گذشته «پزشکِ برای همهٔ اهداف[1]» مانند ابن‌سینا، درمان بیماری را با حداقل ابزار و بر اساس دانش، تجربیات و برخی سنت‌های محلی انجام می‌داد؛ اما اکنون اتاق‌های عمل مجهزی وجود دارند که از دستگاه‌های متعددی برای انجام و نظارت بر درمان و جراحی استفاده می کنند.

علی‌رغم تلاش‌های انجام‌شده برای مداخلهٔ دانش رباتیک، هوش مصنوعی و پردازش تصویر در فرآیندهای درمان و جراحی، هم‌چنان این پزشکان و تیم‌های جراحی هستند که از دانش و تجربیات خود در زمینه‌های پزشکی بهترین استفاده را انجام داده و روند درمان و جراحی را هدایت می‌کنند. در آیندهٔ نه‌چندان دور، جراحی، مبتنی بر پردازش خودکار همهٔ داده‌ی موجود در اتاق عمل، به خصوص داده‌ی تصویری، گسترش خواهد یافت. این منجر به تسهیل و بهینه‌سازی جراحی و آموزش آن با استفاده از روش‌های مبتنی بر علم داده و هوش مصنوعی خواهد شد [۲].

بیش از ۱۸ سال پیش، محققان برجسته در زمینهٔ جراحی به کمک رایانه[2]، کارگاه «اتاق عمل آینده» را برگزار کردند. حدود ۱۰۰ متخصص از جمله پزشکان، مهندسان و کارکنان اتاق عمل در کارگاه آموزشی شرکت کردند تا چشم‌انداز اتاق عمل مبتنی بر علم داده در سال ۲۰۲۰ میلادی را تعریف کنند. بسیاری از مشکلات و چالش‌هایی که در سال ۲۰۰۴ میلادی شناسایی شده‌اند، امروز هم مشاهده می‌شوند و متاسفانه این رشته با سرعتی که محققان در آن زمان انتظار داشتند، پیشرفت نکرده است. به‌عنوان نمونه، یکی از مهم‌ترین چالش‌هایی که در سال ۲۰۰۴ میلادی به آن اشاره شد، دشواری جمع‌آوری و ادغام انواع مختلف اطلاعات و دادهٔ مورد نیاز در یک قالب منسجم، و سپس تحلیل کارآمد آن‌هاست. موضوعی که هم‌چنان از سوی مجموعه‌های پزشکی به‌عنوان یک چالش جدی ذکر می‌شود. با این حال، در سال‌های اخیر و با ایجاد انقلابی در روش‌های علم داده، یادگیری عمیق، پردازش تصویر و بینایی ماشین، و افزایش دسترسی به داده‌ی تصویری جراحی، شتاب جدیدی به این حوزه بخشیده شده است [۳].



شکل ۱ - سیر تکاملی روش‌های جراحی

با وجود این که حجم و اثربخشی جراحی‌ها در دهه‌های اخیر به‌طرز چشم‌گیری افزایش یافته است، اما هم‌چنان جراحی یکی از منشأهای اصلی خطاهای درمانی است. بنا به اطلاعات ثبت شده، تخمین زده می‌شود که در حال حاضر سالانه بیش از ۳۰۰ میلیون عمل جراحی انجام می‌شود[3]، که علی‌رغم پیشرفت‌های وسیع در حوزه‌های مرتبط پزشکی و جراحی، ۹ میلیون مورد از این جراحی‌ها باعث عوارض عمده‌ای برای بیماران می‌شوند. سیستم پزشکی می‌تواند با جلوگیری از عوارض ناشی از خطاهای جراحی، سالانه حداقل حدود ۴۲۰۰۰ پذیرش مجدد بیمار را کاهش داده و حداقل ۶۲.۳ میلیون دلار صرفه‌جویی کند [۴]–[۶]. انجام جراحی نیازمند مهارت‌ها و تجربیات انسانی پیچیده‌ای است که فراگیری آن‌ها سال‌ها آموزش، تمرین و ارزیابیِ تحت نظارت می‌طلبد. از طرف دیگر، با ایجاد ابزارهای خودکار تحلیلی می‌توان مدیریت بهتری روی فرآیندهای جراحی داشت. امروزه مشخص شده است که مهارت‌های فنی ضعیف در جراحی خطر عوارض شدید پس از جراحی را افزایش می‌دهند. برای نمونه، مجموعهٔ قابل توجهی از تحقیقات نشان می‌دهد که مهارت فنی ضعیف جراح به‌دلیل عدم آموزش و ارزیابی صحیح و کافی، با پیامدهای نامطلوب شدیدی از جمله مرگ، عمل مجدد و بستری مجدد بیماران همراه بوده است [۷]. هم‌چنین در مطالعه‌ای در مورد ادعاهای قصور پزشکی در ایالات متحدهٔ آمریکا، خطاهای فنی، که منشأ بیش‌تر آن‌ها عدم مهارت کافی در کار با ابزار بوده است، در ۶۷ درصد مواردی که به مصدومیت و معلولیت‌های دائمی یا مرگ بیماران منجر شده‌اند، دخیل بوده‌اند [۸]. بنابراین، استفاده از ابزارها و سامانه‌های هوشمند هدایت تصویری در انجام، آموزش و

[1] Physician for all purposes
[2] Computer Aided Surgery (CAS)
[3] در سال ۲۰۱۲ میلادی، ۳۱۲.۹ میلیون جراحی مهم در سراسر جهان انجام شد که نسبت به سال ۲۰۰۴ میلادی، ۳۸ درصد افزایش داشته است [۱۱۰].

ارزیابی جراحی اهمیت زیادی دارد. با مدل‌سازی و خودکارسازی فرآیندهای جراحی، و آموزش و ارزیابی مهارت آن با استفاده از هوش مصنوعی، می‌توان نقاط قوت و محدودیت‌های جراحی را روشن کرد و یک کمک هوشمند مبتنی بر تصویر را برای جراحی‌ها ارائه کرد. در شکل ۱ نمایی از روند تکاملی جراحی آورده شده است. چنانچه در این شکل ملاحظه می‌شود، جراحی در گذشته با استفاده از داروها و ابزارهای سنتی صورت می‌گرفته است. امروزه، جراحی در اتاق‌های پیشرفته و در حضور دستگاه‌های متعدد صورت می‌پذیرد که هریک از این دستگاه‌ها وظیفه‌ای را دنبال می‌کنند. در آینده و با توسعهٔ سامانه‌های مبتنی بر تصویر و هوش مصنوعی، تمامی دستگاه‌های موجود در اتاق عمل به‌صورت یکپارچه و در ارتباط کامل با یکدیگر فعالیت خواهند کرد.

تاکنون سامانه‌های رباتیکی کمک‌جراح به‌عنوان ابزاری موثر در بهبود عمل‌کرد و مهارت جراحان به طور گسترده‌ای به کار گرفته شده اند. با این حال، این سامانه‌های هنوز قادر به ارائهٔ کمک خودران به جراح نیستند و فقط به تکرار حرکات انجام‌شده توسط جراح و در برخی موارد کاهش لرزش دست جراح محدود می‌شوند. از این رو، تلاش‌هایی با هدف توسعهٔ راه‌های کمک خودکار به جراحان برای کاهش بار کاری آن‌ها در طول مداخلات مورد توجه قرار داده شده است. توانایی انجام برخی وظایف مستقل و تصمیم‌گیری به‌صورت بلادرنگ توسط ربات، مستلزم درک عمیقی از محیطی است که دستگاه رباتیک در آن کار می‌کند. بنابراین، استنباط تصویری از صحنهٔ جراحی مانند تشخیص این که چه عناصری در صحنه هستند و یا این که چه فرآیندهایی در یک زمان خاص یا در طول یک مداخله اتفاق می‌افتد اهمیت دارد. هم‌چنین برای فهمیدن زمان، کمیت و کیفیت مداخه در فرآیند جراحی، لازم است درکی از مهارت جراحی ایجاد شود و سطح مطلوب‌بودن حرکات جراح در زمان جراحی به‌صورت مداوم سنجیده شود. بنابراین، برای گسترش کاربرد دستگاه‌های رباتیکی کمک‌جراح، حضور سامانه‌های هوشمند هدایت تصویری ضروری است. در سال‌های اخیر، با پیشرفت چشم‌گیر دانش هوش مصنوعی و پردازش تصویر، توسعهٔ روش‌های جراحی، به خصوص در کاربردهای مربوط به «جراحی به کمک ابزار رباتیکی[1]»، با استفاده از ابزارهای هدایت مبتنی بر تصویر گسترش پیدا کرده‌اند.

با توجه به این که افراد، ابزارها، اندام‌ها و بافت‌هایی که در صحنهٔ یک جراحی حضور دارند از تنوع زیادی برخوردارند، امروزه به‌جای مدل‌سازی صریح بر اساس روش های کلاسیک، استفاده از روش‌های پردازش تصویر مبتنی بر یادگیری عمیق بسیار مرسوم شده است. آنتبی و همکاران [۹]، مطالعهٔ جالبی از کاربرد روش‌های یادگیری عمیق و بینایی ماشین در جراحی لاپاراسکوپی ارائه کرده‌اند. هدف آن‌ها، آشنایی پزشکان با این روش جدید بوده است و بر اساس یافته‌های ایشان و گزارش ارزش بالینی آثار، به نتایج امیدبخشی دست پیدا کرده‌اند.

روش‌های مبتنی بر یادگیری عمیق نیازمند حجم مناسبی از داده هستند. امروزه، جمع‌آوری حجم زیادی از داده با همکاری مشترک جراحان و مهندسان میسر شده است. گسترش سامانه‌های خودکار هدایت جراحی و آموزش جراحی، بر مبنای دو نوع دادهٔ فیزیکی-حرکتی و تصویری-ویدیویی صورت می‌پذیرد. جمع‌آوری داده‌ی فیزیکی-حرکتی به دلیل نیازمندی به حس‌گر و تجهیزات جانبی خاص، پُرهزینه است. یک مطالعه در این زمینه نشان داده است که استفاده از تجهیزات ضبط‌کنندهٔ دادهٔ حرکتی برای ارزیابی مهارت جراحان کارآموز جراحی آب‌مروارید، در مرحلهٔ اولیه نیازمند چهار حس‌گر با هزینهٔ حداقل ۲۰۰۰۰ دلار است [۱۰]. در مقابل، دسترسی به دادهٔ تصویری-ویدیویی به‌دلیل حضور دوربین جراحی در بیش‌ترِ اتاق‌های عمل، کم هزینه‌تر است. هم‌چنین اطلاعات حرکتی نیز می‌توانند از طریق ویدیو و روش‌های پردازش تصویر و بینایی ماشین استخراج شوند که در این صورت، این نوع داده هم به نوعی دادهٔ مرتبط با ویدیو محسوب می‌گردند. با دسترس قرار گرفتن دادهٔ تصویری-ویدیویی، پردازش تصویر و بینایی ماشین می‌توانند در حوزه‌های مختلف جراحی و آموزش آن مداخله کرده و بخشی از فرآیندهای مربوطه را خودکار کنند.

وظایف مختلفی در پردازش تصویر و بیایی ماشین با هدف ساختن قسمت‌های مختلف یک سامانهٔ هوشمند هدایت تصویری جراحی گسترش پیدا کرده‌اند. از این میان می‌توان به وظایف طبقه‌بندی، تشخیص[2] و بخش‌بندی[3] نوع و محل ابزار، بافت‌ها و نواحی مهم جراحی، تشخیص فاز جراحی، تشخیص ژست در جراحی و ارزیابی مهارت جراحی اشاره کرد. تمام این وظایف قطعه‌هایی از یک پازل هستند که به هدف خودکارسازی کامل فرآیند جراحی و آموزش آن از طریق سامانه‌های هدایت تصویری کمک می‌کنند.

در این مقاله، اثرگذارترین کاربردهای پردازش تصویر و بینایی ماشین در جراحی و آموزش آن مورد بررسی قرار می‌گیرد. در بخش اول این مقاله تعدادی مجموعه‌دادهٔ تصویری مهم مربوط به جراحی معرفی می‌شوند. در بخش دوم، به معرفی حوزهٔ تحلیل مکانی تصاویر جراحی پرداخته می‌شود که شامل بخش‌هایی هم‌چون طبقه‌بندی، تشخیص، ردیابی و بخش‌بندی اجزای مهم جراحی است. این قسمت برای فهم عمیق صحنهٔ جراحی اهمیت دارد. در بخش سوم، به معرفی پژوهش‌های مربوط به کاربرد پردازش تصویر و بینایی ماشین در تشخیص مراحل، فعالیت‌ها و ژست‌های جراحی پرداخته شده است. این بخش به فهم بهتر فرآیند جراحی و تحلیل کارآمد آن کمک می‌کند. در بخش چهارم، روش‌های مبتنی بر تصویر در ارزیابی مهارت جراحی معرفی شده است. این بخش به ایجاد درک سامانه‌های هوشمند از آن‌چه در یک دسته‌بندی کلی جراحی «خوب» و «بد» شناخته می‌شود کمک می‌کند. در بخش پنجم، جمع‌بندی و افق آیندهٔ پردازش تصویر و بینایی ماشین در حوزهٔ جراحی بیان می‌شوند.

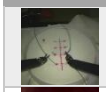
[1] Robot Assisted Surgery (RAS)
[2] Detection
[3] Segmentation

۲- مجموعه‌دادهٔ تصویری-ویدیویی جراحی

امروزه آماده‌سازی و درک داده یکی از مهم‌ترین و زمان‌برترین وظایفی است که در یک پروژهٔ بینایی ماشین و یادگیری عمیق انجام می‌شود. نظرسنجی‌ها نشان می‌دهند که اکثر دانشمندان علم داده و توسعه‌دهندگان هوش مصنوعی نزدیک به ۷۰ درصد از زمان خود را صرف تجزیه‌وتحلیل و ایجاد مجموعه‌داده کرده و زمان باقی‌مانده را صرف فرآیندهای دیگری مانند انتخاب مدل، آموزش، آزمایش و توسعه روش‌ها می‌کنند [۱۱] . در سال‌های گذشته، تلاش‌های زیادی برای ایجاد مجموعه‌دادهٔ عمومی و گسترده از روش‌های جراحی که توسط متخصصان حاشیه‌نویسی شده‌اند، صورت گرفته است.

مجموعه‌دادهٔ JIGSAWS[1] [۱۲]، شامل مجموعه‌ای از دادهٔ سینماتیکی و ویدیویی است که از فرآیندهای سادهٔ جراحی و توسط سامانهٔ جراحی رباتیک داوینچی[2] جمع‌آوری شده است. این فرآیندها شامل سه وظیفهٔ ابتدایی جراحی در فضایی شبیه‌سازی‌شده[3] است و عبارتند از: عبور سوزن[4]، بخیه‌زنی[5] و گره‌زنی[6] که در ستون نمونهٔ سطر اول جدول ۱ نشان داده شده است.

جدول ۱ – معرفی مهم‌ترین مجموعه‌دادهٔ تصویری جراحی.

نام مجموعه‌داده	نوع جراحی	انواع و تعداد داده	انواع حاشیه‌نویسی‌ها	نمونه
JIGSAWS [۱۲]	بخیه‌زنی (محیط غیرواقعی)	۱۰۳ دادهٔ ویدیویی و سینماتیکی	مهارت، مراحل، ژست	
Cholec80 [۱۶]	کوله‌سیستکتومی لاپاروسکوپی	۸۰ دادهٔ ویدیویی	مراحل، تشخیص و بخش‌بندی	
SurgicalActions160	لاپاراسکوپی زنان	۱۶۰ ویدیو	فعالیت (Action)	
GLENDA [۲۱]	لاپاراسکوپی زنان	۲۵۰۰۰ تصویر از ۴۰۰ جراحی	تشخیص آندومتریوز	
LapSig300 [۲۲]	لاپاراسکوپی کولورکتال	۳۰۰ ویدیو	مراحل، فعالیت و مکان ابزار	
Cataract-21 [۲۳]	آب‌مروارید	۲۱ ویدیو	مراحل	
Cataract-101 [۲۴]	آب‌مروارید	۱۰۱ ویدیو	مهارت (۲ سطح، براساس تجربه)	
CaDIS [۲۵]	آب‌مروارید	۴۶۷۰ تصویر از ۲۵ جراحی	بخش‌بندی ابزار و بافت‌ها	
ARAS-Farabi [۲۶], [۲۷]	آب‌مروارید (کپسولرکسیس)	بیش از ۲۰۰۰۰ تصویر از ۱۰۰ ویدیو	تشخیص، بخش‌بندی، ردیابی ابزار و بافت، و مهارت (نمرهٔ ۰ تا ۵ در ۹ شاخصه)	

برای تشکیل مجموعه‌دادهٔ JIGSAWS از ۸ جراح راست‌دست استفاده شده است که هریک میزان تجربهٔ متفاوتی از کار با ابزار جراحی رباتیک داشته‌اند. جراح‌های گروه ماهر بیش از ۱۰۰ ساعت، جراح‌های گروه تازه‌کار کم‌تر از ۱۰ ساعت و جراح‌های گروه متوسط بین ۱۰ تا ۱۰۰ ساعت تجربهٔ جراحی رباتیک داشته‌اند. همهٔ افراد هریک از سه وظیفهٔ تعیین‌شدهٔ جراحی را برای پنج بار تکرار کردند که هربار انجام این سه‌وظیفه به‌طور متوسط حدود دو دقیقه زمان برده است. اطلاعات سینماتیکی مجموعه‌دادهٔ JIGSAWS هم شامل ۱۹ متغیر سینماتیکی است که به ۷۶ بخش تقسیم شده است که در فرکانس ۳۰ هرتز جمع‌آوری و ضبط شده است. دادهٔ ویدویی مجموعه‌دادهٔ JIGSAWS از هر دو دوربین آندوسکوپیک سامانهٔ جراحی داوینچی و با وضوح ۶۴۰ در ۴۸۰ پیکسل ضبط و ذخیره شده‌اند. این دادهٔ ویدیویی، با دادهٔ سینماتیکی همگام شده‌اند. هم‌چنین مجموعه‌دادهٔ JIGSAWS با مشورت پزشکان متخصص، و در نهایت با تأکید بر ساعات تجربهٔ جراحی جراحان، سه فعالیت اصلی خود را در سه سطح مهارتی، نمره‌دهی استاندارد و چند ژست جراحی حاشیه‌نویسی کرده است. لازم به ذکر است که مجموعه‌دادهٔ ATLAS Dione [۱۳]، UCL dVRK [۱۴] و FlapNet10 [۱۵] هم با استفاده از ربات جراح داوینچی جمع‌آوری شده‌اند که هر یک شامل تعدادی ویدیو از فرآیندهای جراحی با استفاده از ابزار رباتیک هستند.

[1] JHU-ISI Gesture and Skill Assessment Working Set
[2] da Vinci Surgical System (dVSS)
[3] لفظ «شبیه‌سازی‌شده» در این‌جا به معنای استفاده از ابزارهای واقعیت افزوده نیست؛ بلکه، به این معناست که JIGSAWS به‌جای داده‌های جراحی واقعی، از داده‌های مدل‌های رومیزی استفاده کرده است. این موضوع با دقت در ستون نمونهٔ سطر اول جدول ۱ مشخص شده است.
[4] Needle-Passing (NP)
[5] Suturing (SU)
[6] Knot-Tying (KT)

مجموعه‌دادهٔ Cholec80 [۱۶]، شامل ۸۰ ویدیو از کوله‌سیستکتومی لاپاروسکوپی است که توسط ۱۳ متخصص بیمارستان دانشگاه استراسبورگ انجام شده و با سرعت ۲۵ قاب تصویری بر ثانیه ضبط شده است. این مجموعه‌داده شامل حاشیه‌نویسی ابزار با نرخ ۱ قاب تصویری بر ثانیه بوده و مناسب فعالیت‌های مربوط به آشکارسازی شئ است. در کاری دیگر [۱۷] حاشیه‌نویسی مراحل جراحی نیز به این مجموعه‌داده اضافه شد. مجموعه‌دادهٔ mc2cai16-tool هم بعدتر روی همین داده و با ایجاد حاشیه‌نویسی مکانی برای ابزار و آماده‌سازی آن برای فعالیت‌های ردیابی و ارزیابی مهارت جراحی توسعه داده شد [۱۸].

مجموعه‌دادهٔ SurgicalActions160 [۱۹]، شامل کلیپ‌های ویدیویی کوتاهی است که نشان‌دهندهٔ ۱۶ عمل معمول در لاپاراسکوپی زنان است که از جراحی‌های مختلف گردآوری شده است. برای هر کلاس عمل دقیقا ۱۰ نمونه کلیپ وجود دارد. مجموعه‌دادهٔ LapGyn4 [۲۰][1] هم شامل بیش از ۵۵ هزار تصویر از بیش از ۵۰۰ جراحی لاپاراسکوپی و بخش‌های اقدامات جراحی، ساختارهای آناتومی، اقدامات روی آناتوم، شمارش ابزار و تجزیه‌وتحلیل محتوای خودکار جمع‌آوری شده است.

مجموعه‌دادهٔ GLENDA [۲۱][2]، شامل بیش از ۲۵۰۰۰ تصویر است که از بیش از ۴۰۰ جراحی لاپاراسکوپی زنان گرفته شده است و به طور هدفمند ایجاد شده است تا برای انواع مسائل مربوط به تحلیل محتوای خودکار در تشخیص آندومتریوز [3] مورد استفاده قرار گیرد. این مجموعه‌داده شامل حدود ۱۲۰۰۰ قاب تصویری است که اندومتریوز را با شدت‌های مختلف نشان می‌دهد، و هم‌چنین حدود ۱۳۰۰۰ قاب تصویری که آندومتریوز را نشان نمی‌دهد. بسیاری از قاب‌های تصویری مبتلا به اندومتریوز حاوی حاشیه‌نویسی‌های تخصصیِ مبتنی بر ناحیه و زمان هستند.

مجموعه‌دادهٔ LapSig300 [۲۲]، اولین مجموعه‌داده در مقیاس بزرگ از جراحی لاپاراسکوپی کولورکتال است. این مجموعه‌داده، شامل ۳۰۰ ویدیوی به دست آمده از جراحی‌های ۱۹ مرکز آندوسکوپی بزرگ در ژاپن است. این جراحی‌ها توسط جراحان مختلف در طول ۱۰ سال انجام شده‌اند. قاب‌های تصویری این مجموعه‌داده دارای حاشیه‌نویسی‌هایی از مرحله، فعالیت و مکان ابزار هستند.

مجموعه‌دادهٔ Cataract-21 [۲۳]، شامل ۲۱ فیلم ضبط‌شده از جراحی‌های آب‌مروارید است. این مجموعه‌داده به یک بخش آموزشی شامل ۱۷ ویدیو و یک بخش اعتبارسنجی شامل ۴ ویدیو تقسیم می‌شود. برای هر ویدیو یک فایل با حاشیه‌نویسی‌های واقعی ارائه می‌شود، که هر شماره قاب تصویری را به یکی از ده کلاس (مراحل جراحی) تعیین‌شده در جراحی آب‌مروارید پیوند می‌دهد. مجموعه‌دادهٔ تصویری Cataract-101 [۲۴]، شامل ویدیوهای ضبط‌شده از ۱۰۱ عمل آب‌مروارید است که توسط چهار جراح مختلف در بخش چشم‌پزشکی و بینایی‌سنجی در بزرگترین بیمارستان دولتی در کلاگنفورت اتریش انجام شده است. این ویدیوها در یک دورهٔ نه‌ماهه جمع‌آوری شده و در آن ده مرحله و فاز از جراحی آب‌مروارید توسط یک جراح ارشد چشم حاشیه‌نویسی شده است. هم‌چنین این چهار جراح با توجه به تعداد کل جراحی‌های انجام‌شده و موقعیت شغلی‌شان در دو سطح مهارتی مختلف گروه‌بندی شده‌اند. دو نفر از آن‌ها دستیار جراح با تجربهٔ متوسط (سطح ۱) و دو نفر دیگر جراحان ارشد با تجربهٔ بالا (سطح ۲) هستند.

مجموعه‌دادهٔ CaDIS [۴] [۲۵]، مجموعه‌دادهٔ دیگری در زمینهٔ جراحی آب‌مروارید است که شامل ۴۶۷۰ تصویر است که از ۲۵ ویدیوی جراحی نمونه‌برداری شده است. در هر پیکسل هر تصویر این مجموعه‌داده، ابزارها و بافت‌های تعیین‌شده بخش‌بندی و برچسب‌گذاری شده است.

گروه رباتیک ارس در دانشگاه خواجه نصیرالدین طوسی نیز با همکاری بیمارستان چشم‌پزشکی فارابی تعدادی مجموعه‌داده تصویری دارای حاشیه‌نویسی‌های مکانی و مهارتی از جراحی کپسولرکسیس ارائه کرده است [۲۶]، [۲۷]. این مجموعه‌داده می‌تواند برای کاربردهای تشخیص، بخش‌بندی و ردیابی ابزار و نواحی مهم جراحی، و ارزیابی مهارت جراحی به کار رود. در یکی از نسخه‌های این مجموعه‌داده، حاشیه‌نویسی‌های مهارتی مربوط به ۱۲۵ ویدیو اضافه شده است. این ویدیوها مربوط به یکی از سرنوشت‌سازترین فرآیندهای جراحی آب‌مروارید با نام کپسولرکسیس هستند. این داده توسط سه جراح متخصص و در نُه شاخصهٔ مخصوصِ این جراحی با امتیاز بین ۰ تا ۵ امتیازدهی شدند و با در نظر گرفتن آستانه‌هایی برای نمرات کلی جراحی، ویدیوها در دو دستهٔ ماهر و تازه‌کار (متوسط) تقسیم‌بندی شده‌اند.

لازم به ذکر است که پژوهش‌های بسیاری در زمینهٔ هوش مصنوعی و جراحی، مجموعه‌دادهٔ مخصوص به خود را گزارش می‌کنند و طبیعتاً تعداد مجموعه‌داده در حوزهٔ جراحی از این مقدار بیش‌تر است. در جدول ۱ شماری از مهم‌ترین مجموعه‌داده‌های معرفی‌شده در این بخش به‌صورت خلاصه معرفی شده‌اند. همان‌طور که مشخص است، در میان مجموعه‌دادهٔ موجود، مجموعه‌دادهٔ ارس-فارابی در تعداد و کیفیت داده و حاشیه‌نویسی‌های آن بسیار جامع و ارزشمند است.گروه رباتیک ارس در حال حاضر با ثبت درخواست در سایت این گروه [5] این مجموعه‌داده را در دسترس قرار می‌دهد.

علاوه بر مجموعه‌دادهٔ ذکرشده در این بخش، چالش‌های مختلفی در سطح جهان با هدف تشکیل و تکمیل مجموعه‌دادهٔ جراحی و پیاده‌سازی مدل‌های هوش مصنوعی روی آن‌ها برگزار می‌شود. از جملهٔ این چالش‌های می‌توان به EndoVis و M2CAI16 اشاره کرد. مجموعه‌دادهٔ برآمده از این چالش‌ها معمولاً با همین نام و سال برگزاری آن، منتشر می‌شوند. مثلاً مجموعه‌دادهٔ EndoVis17 مربوط اجرای این

[1] Gynecologic Laparoscopy Dataset
[2] ITEC Gynecologic Laparoscopy Endometriosis
[3] Endometriosis
[4] Cataract Dataset for Image Segmentation
[5] www.aras.kntu.ac.ir /datasets

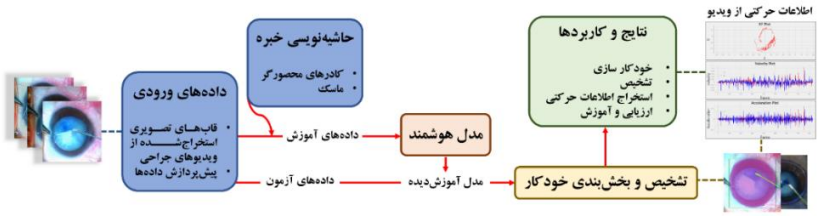
چالش در سال ۲۰۱۷ است. از سال ۲۰۲۳ میلادی، گروه رباتیک ارس هم چالشی با همین هدف در کنفرانس بین‌المللی ایکرام[1] برگزار می‌کند.

۲- تشخیص، ردیابی و بخش‌بندی در تصاویر جراحی

تجزیه و تحلیل تصویر جراحی و تشخیص نواحی (اشیاء) مهم در آن، برای درک سناریوی جراحی و توانایی تحلیل، استدلال و تصمیم‌گیری در فرآیندهای مختلف جراحی، ضروری است. با اجرای فرآیندهای مرتبط با تحلیل تصاویر جراحی، می‌توان روند تغییر اطلاعات مکانی بافت‌ها و ابزارها، و هم‌چنین نحوهٔ تعامل‌های دست جراح در صحنهٔ جراحی را تحلیل کرد. هم‌چنین، اطلاع از نوع ابزارهای جراحی مورد استفاده در یک زمان معین، اطلاعاتی کلیدی از آن‌چه در جراحی در جریان است ارائه می‌دهد. بنابراین، توانایی تحلیل تصویر جراحی و تشخیص ابزار جراحی، بافت‌ها و ساختارهای آناتومی در صحنه، برای آگاهی سامانه‌های جراحی از صحنهٔ جراحی ضروری است. هم‌چنین، تشخیص و بخش‌بندی سایر مواد جراحی، مانند نخ بخیه، اطلاعات مفیدی را برای خودکارسازی فرآیندهای جراحی فراهم می‌کند. الگوریتم‌های پردازش تصویر و بینایی ماشین وظایف مربوط به تشخیص، بخش‌بندی و ردیابی در تصاویر جراحی را به‌صورت خودکار درمی‌آورند. بسته به خروجی مورد انتظار از مدل، روش‌های تشخیص اشیاء در یک تصویر را می‌توان به موارد زیر تقسیم کرد:

- طبقه‌بندی: در برخی کاربردها، حضور یک ابزار خاص در جراحی از اهمیت برخوردار است. مثلاً با طبقه‌بندی در قاب‌های تصویری ویدیوی جراحی‌ای که در آن ابزار کپسولرکسیس حضور پیدا کرده، مرحلهٔ جراحی آب‌مروارید که مربوط به این فرآیند است مشخص می‌شود. این طبقه‌بندی می‌تواند به‌صورت یک یا چندکلاسه انجام شود.
- تشخیص: این الگوریتم‌ها برای مکان‌یابی اشیاء در تصویر و نشان‌دادن آن‌ها با کادرهای محصورگر استفاده می‌شوند. با تعیین محل حضور ابزار جراحی در کادر تصویر، می‌توان وضعیت مهارت‌های حرکتی جراح را به‌صورت کمّی مورد بررسی قرار داد. کادرهای محصورگر معمولاً مستطیل‌شکل هستند؛ بدین دلیل با آن نمی‌توان اشکال یا لبه‌ها را تشخیص داد.
- بخش‌بندی: در این روش هر شئ با ماسک‌های[2] پیکسلی در تصویر مشخص می‌شود. نوع اشیاء را می‌توان به‌صورت دودویی بخش‌بندی کرد به صورتی که هر پیکسل در یک تصویر به عنوان ابزار یا پس‌زمینه برچسب‌گذاری شود. این کار را می‌توان به صورت چندکلاسه هم انجام داد. بخش‌بندی، فهم دقیق‌تری از شکل و لبه‌ها به دست می‌دهد و اطلاعات استخراج‌شده از آن دقت بالاتری خواهند داشت.
- ردیابی: الگوریتم‌های ردیابی کمک می‌کند که در یک ویدیو و در طول زمان اطلاعاتی از ابزارها و بافت‌های حاضر و هم‌چنین موقعیت‌های مکانی آن‌ها داشته باشیم. با داشتن این اطلاعات می‌توان درکی عمیق‌تر از آن‌چه در یک فرآیند جراحی اتفاق افتاده و یا در جریان است به دست آورد.

نمایی از چگونگی و کاربرد وظایف مطرح‌شده در این بخش، در شکل ۲ آورده شده است. همان‌طور که مشاهده می‌شود، این وظایف در نهایت به فهم سامانه‌های کمک‌جراحی از موقعیت ابزارها و بافت‌ها در سناریوی جراحی، استخراج اطلاعات حرکتی و استفاده از این اطلاعات برای تحلیل و مداخله منجر می‌شود



شکل ۲ – فرآیند و کاربرد تشخیص و بخش‌بندی اجزای جراحی.

۲-۱- طبقه‌بندی، تشخیص و ردیابی در تصاویر جراحی

با استخراج محل ابزار در کادر تصویر و نسبت به محیط جراحی می‌توان فرآیند جراحی یا آموزش جراحی را تحلیل و به‌صورت کارآمد و دقیقی در آن مداخله کرد. با گسترش ابزارهای رباتیکی کمک جراحی، وجود یک سامانه که به‌صورت خودکار اطلاعات مکانی ابزارها و نواحی مهم جراحی را به دست آورد از اهمیت بیش‌تری برخوردار می‌شود. از جمله چالش‌های این اقدام می‌توان به مشکلاتی مانند تغییر مقیاس، شباهت ابزارها به یکدیگر و یا به برخی اجزای صحنه مانند بازتاب نور، و پوشانده‌شدن بخشی از نواحی توسط دست جراح اشاره کرد. تاکنون کارهای مختلفی برای خودکارسازی این وظیفه انجام شده است که از روش‌های مبتنی بر پردازش تصویر و بینایی ماشین بهره برده‌اند.

بوگت و همکاران [۲۸]، یک مجموعه‌دادهٔ تشخیص ابزار جراحی و روشی مبتنی بر پردازش تصویر برای تشخیص ابزار و تخمین موقعیت آن‌ها در تصاویر دوبُعدی ارائه کرده‌اند. در ساختار پیشنهادی آن‌ها، ابتدا با استفاده از روش‌های کلاسیک پردازش تصویر مجموعه‌ای از کانال‌های ویژگی یکپارچه از تصویر ورودی استخراج می‌شود. در ادامه ، هر پیکسل فقط بر اساس ظاهر محلی طبقه‌بندی می‌شود. درنهایت یک الگوی خاص

[1] www.icrom.ir
[2] Masks

ابزار روی نقاط تشخیص‌داده‌شده اعمال می‌شود و فرآیند تخمین موقعیت صورت می‌پذیرد. مشکل اصلی این کار، عدم تشخیص قسمت‌هایی از ابزار در تصویر است. برای بهبود نتایج، باید از روش‌های پیشرفتهٔ پردازش تصویر استفاده شود.

در ویدیوهای جراحی لاپاراسکوپی، هر تصویر در یک زمان، معمولاً دارای بیش از یک ابزار است. بنابراین، طبقه‌بندی به‌صورت چندکلاسه انجام می‌شود که در آن هر نمونه می‌تواند به بیش از یک کلاس تعلق داشته باشد. ناکازاوا و همکاران [۳۲]، یک الگوریتم بلادرنگ برای تشخیص سوزن جراحی در ویدیو را با استفاده از یک شبکهٔ عصبی کانولوشنی (CNN) مبتنی بر ناحیه توسعه دادند. در نتیجهٔ این کار، سوزن حتی زمانی که به شدت توسط ابزار ویا رگ‌های خونی در طول آناستوموز میکروواسکولار مسدود شده باشد، به خوبی تشخیص داده می شود. هرچند در برخی از موارد قسمتی از سوزن به درستی تشخیص داده نمی‌شود.

ونگ و همکاران [۲۹]، یک روش طبقه‌بندی چندکلاسه ارائه کردند که دو مدل شبکهٔ کانولوشنیِ VGGNet [۳۰] و GoogLeNet [۳۱] را برای تولید نتیجهٔ نهایی ترکیب می‌کند. در ساختار پیشنهادی آن‌ها، هر شبکه به طور جداگانه آموزش داده می‌شود و از پیش‌بینی هر یک از آن‌ها برای محاسبهٔ طبقه‌بندی نهایی میانگین گرفته می‌شود. محدودیت اصلی این کار در نظر نگرفتن اطلاعات زمانی ویدیوها است. اطلاعات زمانی، در تشخیص ترتیب ورود ابزار و غلبه بر مشکل شباهت زیاد ابزارهای جراحی به یکدیگر حائز اهمیت است.

میشرا و همکاران [۳۲]، پیشنهاد کردند که اطلاعات مکانی-زمانی، با استفاده از یک شبکهٔ عصبی عمیق بازگشتی حافظه طولانی کوتاه مدت (LSTM) در مسألهٔ طبقه‌بندی ابزار در نظر گرفته شود. در مرحلهٔ اول، یک CNN برای تشخیص وجود ابزار در قاب‌های تصویری جداگانه آموزش می‌بیند؛ سپس، ویژگی‌های آموخته‌شده توسط CNN برای یادگیری یک مدل زمانی با استفاده از یک شبکهٔ LSTM استفاده می‌شود. این کار به دقت بالاتری از طبقه‌بندی ابزار جراحی رسیده است.

الحاج و همکاران [۳۳]، یک ابزار نظارتی را در طول جراحی با استفاده از شبکه‌های عصبی کانولوشنی و بازگشتی (RNN) پیشنهاد کرده‌اند. ساختار پیشنهادی آن‌ها شامل چندین CNN است که ویژگی‌های دیداری ویدیوها را استخراج می‌کنند. RNNها هم بر اساس خروجی‌های CNN، توالی زمانی ویژگی‌ها در کل جراحی را تجزیه و تحلیل می‌کنند. با این رویکرد، آن‌ها به عمل‌کردی با دقت ۹۸٪ دست پیدا کرده‌اند.

ساریکایا و همکاران [۱۳] در سال ۲۰۱۷، اولین روی‌کرد مبتنی بر شبکه‌های عصبی عمیق برای تشخیص و شناسایی محل ابزار در ریات‌های کمک‌جراح را پیاده‌سازی کردند. آن‌ها از یک شبکهٔ پیشنهاد ناحیه (RPN) [1] در ترکیب با یک شبکهٔ کانولوشنی چندوجهی [2] برای مشخص‌کردن محل اشیاء، و یک R-CNN سریع برای تشخیص اشیاء استفاده کردند. در این کار آن‌ها مجموعه‌داده‌ای با نام ATLAS Dione معرفی کردند که اولین مجموعه‌دادهٔ عمومی از ویدیوهای جراحی به کمک ربات و با حاشیه‌نویسی ابزار بوده است. دو [۳۴] و کولنی [۳۵] با مدل‌سازی ابزارها به‌صورت مجموعه‌ای از مفاصل و اتصالات بین آن‌ها، از یک شبکهٔ عصبی کانولوشنی برای شناسایی مفاصل مشترک، و از این خروجی برای تخمین موقعیت ابزارها استفاده کرده‌اند.

چن و همکاران [۳۶]، با استفاده از یک CNN به‌همراه آشکارساز قطعه‌ای خط [3] خطوط ابزار و ویژگی‌های آن‌ها را شناسایی کرده و به ردیابی بلادرنگ تصویر به تصویر ابزار جراحی در ویدوی جراحی پرداختند. ژاو و همکاران [۳۷] یک CNN آبشاری برای شناسایی و تشخیص محل ابزار جراحی ربایتک پیشنهاد کردند. این شبکه نقشهٔ حرارتی ناحیهٔ نوک ابزار را خروجی می‌دهد. این کار از یک شبکهٔ VGG-16 اصلاح‌شده برای رگرسیون کادر محصورگر روی این نقشه‌های حرارتی استفاده کرده است. لیو و همکاران [۳۸]، با توسعهٔ این کار و استفاده از یک CNN بدون لنگر [4] ابزارهای جراحی را به‌عنوان یک نقطه مدل‌سازی کردند. این کار موفق به بهبود سرعت و دقت در زمینهٔ تشخیص ابزار جراحی شد.

ونگ و همکاران [۳۹]، شبکهٔ YOLOv7x را برای کاربرد تشخیص و شمارش ابزار جراحی بهبود و توسعه دادند. آن‌ها یک ماژول با نام RepLK Block معرفی کردند که می‌تواند علاوه بر افزایش میدان دید، ویژگی‌های دیداری بیش‌تر و بهتری را برای این کاربرد استخراج کند.

در میان روش‌های مورد استفاده برای انجام طبقه‌بندی و تشخیص در تصاویر جراحی، CNNها و نسخه‌های مختلفی از مدل‌هایی مانند YOLO [5] به‌دلیل توانایی در تشخیص و تعیین موقعیت چندین ابزار، مرسوم‌ترین معماری بوده‌اند. مقالاتی که در این حوزه معرفی شده‌اند توانسته‌اند روی مجموعه‌دادهٔ مختلف عمومی و خصوصی به دقت‌هایی بین ۸۵ تا ۱۰۰ درصد دست پیدا کنند.

در زمینهٔ ردیابی ابزار جراحی هم کارهای مهمی صورت گرفته است. نویه و همکاران [۴۰]، یک روی‌کرد مبتنی بر CNN و LSTM پیشنهاد کرده‌اند. آن‌ها ضمن مدل‌سازی وابستگی‌های زمانی در حرکت ابزارهای جراحی و استخراج اطلاعات مکانی، به ردیابی ابزارهای جراحی پرداخته‌اند. آن‌ها راه‌کاری هم برای دشواری حاشیه‌نویسی دستی مجموعه‌داده ارائه کرده‌اند. اسلام و همکاران [۴۱]، یک مدل یادگیری چندوظیفه‌ای پیشنهاد کردند که دارای یک رمزگذار و دو رمزگشاست. آن‌ها هم‌چنین یک تابع اتلاف مخصوص برای بهبود ایجاد نواحی دیداری متمایز معرفی کردند. دو و همکاران [۴۲]، یک چهارچوب ردیابی برای حل چالش تشخیص مقیاس و پس‌زمینه در ردیابی ابزار جراحی پیشنهاد

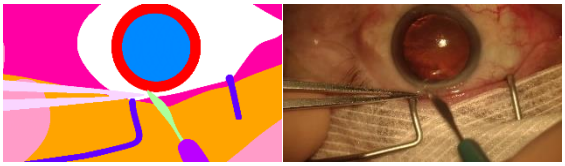
[1] Region Proposal Network (RPN)
[2] Multimodal Convolutional Network
[3] Line Segment Detector (LSD)
[4] Anchor
[5] You only look once (YOLO)

کردند که ضمن توجه به اطلاعات رنگی، به ردیاب اجازه می‌دهد تا تغییرات مقیاس ناگهانی بین قاب‌های تصویری ویدیو را مدیریت کند.

در گروه رباتیک ارس هم فعالیت‌هایی در زمینهٔ خودکارسازی فرآیند طبقه‌بندی و تشخیص ابزار جراحی کپسولرکسیس و ناحیهٔ اصلی جراحی (عنبیه و مردمک چشم) صورت گرفته است. در یکی از کارهای منتشرشده در این زمینه [۲۶] ، ضمن معرفی یک مجموعه‌دادهٔ جدید برای تشخیص ابزار و نواحی مهم جراحی کپسولرکسیس، نتیجهٔ جدیدترین الگوریتم‌های بینایی ماشین در این کاربرد مقایسه شده است. در این کار، یک چهارچوب مبتنی بر یادگیری عمیق برای فعالیت طبقه‌بندی و تشخیص اشیاء معرفی شده است که قابلیت استفاده از مجموعه‌داده و الگوریتم‌های مختلف بینایی ماشین را به سادگی فراهم می‌کند. این چهارچوب از طریق گیت‌هاب[1] در دسترس است. هم‌چنین درکاری دیگر [۲۷]، گروه رباتیک ارس به معرفی و پیاده‌سازی ساختارهای مربوط به ردیابی ابزار و نواحی اصلی جراحی کپسولرکسیس پرداخته است.

۲-۲- بخش‌بندی در تصاویر جراحی

با اجرای الگوریتم‌های تشخیص در جراحی، اطلاعات مکانی ابزار به یک کادر مستطیل‌شکل محدود خواهد شد. بخش‌بندی کمک می‌کند تا اطلاعات دقیق‌تری از مکان، شکل‌ها و لبه‌های ابزارها و نواحی جراحی استخراج شود. به کار بردن این اطلاعات علی‌رغم پیچیدگی بیش‌تر، کارآمدی ابزارهای هدایت و تحلیل تصویری جراحی را بالاتر خواهد برد. نمونه‌ای این فرآیند در شکل ۳ نشان داده شده است. همان‌طور که در این شکل مشخص است، اجزای مختلف حاضر در این تصویر از جراحی چشم، بخش‌بندی شده‌اند. این بخش‌بندی به ابزار خودکار هدایت تصویری کمک خواهد کرد تا اطلاعات بیش‌تری از صحنهٔ جراحی داشته باشد. تاکنون تلاش‌های مختلفی در این زمینه انجام شده که در ترکیب آن با وظیفهٔ ردیابی می‌توان ابزارهای بخش‌بندی ویدیویی یا ردیابی مبتنی بر بخش‌بندی را توسعه داد.



شکل ۳ – نمونه‌ای از بخش‌بندی در جراحی.

لایانا و همکاران [۴۳]، یک روش جدید برای بخش‌بندی و ردیابی بلادرنگ ابزار جراحی پیشنهاد کرند که از یک CNN یکپارچه برای ایجاد وابستگی متقابل بین دو وظیفهٔ محل و بخش‌بندی بهره می‌برد. گارسیا و همکاران [۴۴]، این کار به با پیشنهاد معماری سبکی به نام ToolNet برای ویدیوهای اندوسکوپی توسعه دادند. معماری پیشنهادی آن‌ها پارامترهای کم‌تری دارد و به حافظهٔ کم‌تری هم نیاز دارد. درنتیجه، امکان بخش‌بندی و ردیابی به‌صورت بلادرنگ به‌صورت بهتری فراهم می‌شود. روش آن‌ها از استخراج ویژگی‌های چندمقیاسی استفاده می‌کند. میلتاری و همکاران [۴۲]، یک معماری مبتنی بر رمزگذار–رمزگشا پیشنهاد کرده‌اند که در آن رمزگذار یک شبکه بسیار عمیق مبتنی بر یادگیری باقی‌مانده است و رمزگشا یک شبکهٔ مبتنی بر CNN-LSTM است. این مدل برای بخش‌بندی باینری و چندکلاسه به دقت بیش از ۹۲ درصد دست یافته است.آن‌ها این روش را روی مجموعه‌دادهٔ چالش EndoVis 2015 و برای بخش‌بندی تصاویر جراحی پیاده‌سازی کردند.

شوتس و همکاران [۴۵]، چهار معماری یادگیری عمیق مختلف را برای بخش‌بندی ابزارهای جراحی پیشنهاد کردند: U-Net [۴۶] اصلاح‌شده، دو حالت اصلاح‌شده از TernausNet [۴۷] و یک حالت اصلاح‌شده از LinkNet [۴۸]. به عنوان یک پیشرفت نسبت به U-Net، آن‌ها از شبکه‌های مشابه با رمزگذارهای از پیش آموزش دیده استفاده کردند. TernausNet یک معماری U-Net–مانند است که از شبکه‌های نسبتاً سادهٔ VGG11 یا VGG16 که از پیش آموزش دیده اند به‌عنوان رمزگذار استفاده می‌کند. مدل LinkNet از یک رمزگذار مبتنی بر معماری ResNet استفاده می‌کند. در این کار از ResNet34 پیش آموزش دیده استفاده شده است. بخش رمزگشای این شبکه هم از چندین بلوک رمزگشا تشکیل شده است که به بلوک رمزگذار مربوطه متصل می‌شوند. هر بلوک رمزگشا شامل عملیات کانولشنی ۱ × ۱ است که تعداد فیلترها را به میزان ۴ عدد کاهش می‌دهد. هم‌چنین از نرمال‌سازی دسته و ترانهادهٔ کانولوشن استفاده شده است تا از نگاشت ویژگی نمونه‌برداری شود. اگرچه معماری LinkNet-34 به‌دلیل استفاده از رمزگذار سبک‌تر از نظر کارایی محاسباتی، سریعترین مدل بوده است، آن‌ها بهترین عملکرد برای بخش‌بندی باینری و چندکلاسه را با استفاده از معماری‌های TernausNet اصلاح‌شده به دست آوردند. پاخوموف و نواب [۴۹]، از یک شبکهٔ تماماًکانولوشنی با بدنهٔ ResNet-18 استفاده کردند و به نتایج بهتری دست یافتند. شبکهٔ پیشنهادی آن‌ها، امکان بخش‌بندی بلادرنگ روی تصاویر با کیفیت بالای جراحی را تا سرعت ۱۲۴ قاب تصویری بر ثانیه فراهم می‌کند.

حسن و لینت [۵۰]، یک معماری U-Net اصلاح‌شده، به نام U-NetPlus، برای بخش‌بندی ابزار جراحی پیشنهاد کردند که از یک مدل از پیش آموزش دیده به عنوان رمزگذار با نرمال‌سازی دسته استفاده می‌کند. در بخش رمزگشا، آن‌ها لایهٔ Deconvolution را با یک لایهٔ Upsampling جایگزین کردند که از درون‌یابی نزدیکترین همسایه و به دنبال آن دو لایهٔ کانولوشنی استفاده می‌کند. نی و همکاران [۵۱]، بخش‌بندی ابزارهای رباتیک جراحی را با استفاده از معماری رمزگذار–رمزگشا مورد بررسی قرار دادند. آن‌ها از معماری سبک وزن MobileNetV2 به عنوان رمزگذار استفاده کردند و از رمزگشای دارای توجه[2] سبک سفارشی برای بازیابی جزئیات مکان استفاده کردند. اسلام و همکاران [۵۲]، با استفاده شبکهٔ ResNet-18 به‌همراه LSTM، کار آن‌ها را با سرعتی مشابه؛ اما عملکرد بخش‌بندی بهتر توسعه دادند.

[1] www.github.com/aras-labs/ARAS-DeepLearning-FW
[2] Attention

گرچه بخش‌بندی ابزارهای جراحی توجه بسیاری از محققان هوش مصنوعی را به خود جلب کرده است، جراحی یک فعالیت پیچیده است که شامل بسیاری از اشیاء دیگر است که تشخیص آن‌ها برای تجزیه و تحلیل عمیق و درک یک صحنهٔ جراحی ضروری است. این اجزای دیگر می‌توانند نخ بخیه یا ساختارهای آناتومی بدن باشند. برای مثال در جراحی آب‌مروارید، بخش‌بندی مردمک و عنبیهٔ چشم هم به‌دلیل اهمیتی که تعامل ابزار جراحی با آن ناحیه دارد، از اهمیت بالایی برخوردار است. یا مثلاً در جراحی لاپاراسکوپی بخش‌بندی کبد به دلیل این که این اندام در طول جراحی دچار تغییر شکل‌های زیادی می‌شود و معمولاً توسط سایر اندام‌ها هم‌پوشانی دارد چالش‌برانگیز است.

نذیر و همکاران [۵۳]، روشی را برای جستجوی بخشی از نمای کبد از نمای کامل مربوطه در زمان اجرا پیشنهاد کرده‌اند. آن‌ها ضمن استفاده از یک شبکهٔ کانولوشنی پیشنهادی، از یک هرم تصویری با استفاده از اندازه‌های مختلف از یک تصویر ورودی با نمای کامل استفاده کردند تا بر مشکل تغییر مقیاس غلبه کنند. فو و همکاران [۵۴]، اثر افزودن دادهٔ برچسب‌دار یا بدون برچسب بیش‌تر را برای بهبود وظایف بخش‌بندی، به ویژه در بخش‌بندی کبد، مطالعه کردند و به این نتیجه رسیدند که اگرچه افزودن دادهٔ برچسب‌دار بیش‌تر، بخش‌بندی را بهبود می‌بخشد؛ اما، استفاده از دادهٔ بدون برچسب بیش‌تر در یک یادگیری نیمه‌نظارتی می‌تواند به سطح قابل مقایسه‌ای از دقت در بخش‌بندی دست یابد.

در سال ۲۰۲۳، با معرفی ابزار همه‌کارهٔ شرکت متا[1] برای بخش‌بندی که SAM[2] [۵۵] نامیده شده است، این امکان فراهم آمده تا حاشیه‌نویسی‌های موردنیاز برای بخش‌بندی آسان‌تر و به‌صورت نیمه‌نظارتی و تقریباً خودکار انجام شود. این مدل، از ترنسفورمرها[3] استفاده می‌کند و روی مجموعه‌داده‌ای بسیار بزرگ آموزش دیده است و قادر است بدون نیاز به آموزش اجزای مختلف تصاویر مختلف ازجمله یک تصویر جراحی را بخش‌بندی کند. با تنظیم دقیق مدل‌های مختلف برآمده از SAM، می‌توان دقت آن را در بخش‌بندی تصاویر جراحی ارتقاء داد. ونگ و همکاران [۵۶]، ظرفیت‌های SAM را در ربات‌های جراح و ابزارهای هدایت تصویری جراحی بررسی کرده‌اند. پارانجاپه و همکاران [۵۷] هم ساختار SAM را برای کارآمدی بیش‌تر در صحنه‌های جراحی توسعه دادند.

جدول ۲ – شماری از جدیدترین پژوهش‌های مهم در زمینهٔ تشخیص، ردیابی و بخش‌بندی در جراحی

هدف	سال و مرجع	مدل بینایی ماشین	مجموعه‌داده	نتایج
تشخیص نوک ابزار	۲۰۲۱ [۵۸]	RetinaNet, YOLO-v2	اختصاصی با ۲۳۱۰ تصویر از ۹ ویدیو	F1 Score = ۸۴.۶%
تشخیص شئ در جراحی	۲۰۲۱ [۲۶]	YOLO-v4, Faster-RCNN, SSD, MobileNet, ResNet, VGG	ARAS-Farabi (اختصاصی)	mAP = ۹۴.۴-۹۹.۲%
ردیابی ابزار	۲۰۲۱ [۵۹]	TernauesNet, MobileNet, ShuffleNet	اختصاصی با ۱۸۴۶ تصویر	Accuracy = ۸۵.۸۷%
تشخیص و بخش‌بندی ابزار	۲۰۲۱ [۶۰]	YOLACT	اختصاصی با ۵۳۱۹ تصویر از ۷۰ ویدیو	Accuracy = ۹۱.۲% Dice = ۴۸.۲%
بخش‌بندی ابزار	۲۰۲۱ [۶۱]	CycleGAN, U-Net	EndoVis17	Dice = ۹۲.۸% IoU = ۸۴.۷%
بخش‌بندی بلادرنگ ابزار	۲۰۲۱ [۶۲]	DMNet, LSTM	EndoVis17 و EndoVis18	mDice = ۷۷.۵۳-۶۱.۰۳% mIoU = ۶۷.۵-۵۳.۸۹%
کنترل موقعیت ربات با بخش‌بندی	۲۰۲۱ [۶۳]	YOLO-v3, ResNet	EndoVis17	IoU = ۸۶.۶%
تشخیص شئ در جراحی	۲۰۲۲ [۶۴]	YOLO-v4, Faster-RCNN, MobileNet, EfficientNet	اختصاصی با ۸۷۰ تصویر از ۱۹۶.۵۵ دقیقه ویدیو	mAP = ۲۲.۲-۳۳.۶% F1 Score = ۷۵.۸۶-۹۳.۵%
ردیابی شئ در جراحی	۲۰۲۲ [۲۷]	SiamBAN, GradNet	ARAS-Farabi (اختصاصی)	IoU = ۶۹.۸۹%
بخش‌بندی ابزار	۲۰۲۲ [۶۵]	SurgiNet, MobileNet-v2	EndoVis17	Mean IoU = ۶۳.۳-۸۹.۱۴%
بخش‌بندی ابزار	۲۰۲۲ [۶۶]	Mask R-CNN, CNN, Swin Transformer	EndoVis17	mIoU = ۵۸.۷۳-۷۴.۰۸%
تشخیص و ردیابی سه‌بُعدی ابزار	۲۰۲۳ [۶۷]	YOLO-v5	اختصاصی با ۱۰۳۴۳۷ تصویر از ۳۴ ویدیو	mAP = ۷۹.۳%
بخش‌بندی ابزار	۲۰۲۳ [۵۷]	Adaptive SAM	EndoVis17 و EndoVis18	گزارش بهبود نسبی نتایج نسبت به مدل عام SAM.

[1] Meta
[2] Segment Anything Model (SAM)
[3] Transformers

در جدول ۲، شماری از مهم‌ترین پژوهش‌های منتشرشدهٔ این زمینه در سال‌های اخیر آورده شده است. برخی پژوهش‌ها مجموعه‌دادهٔ «اختصاصی» ارائه کرده‌اندکه با رجوع به آن‌ها و مکاتبه با نویسندگان، قابل دریافت هستند.

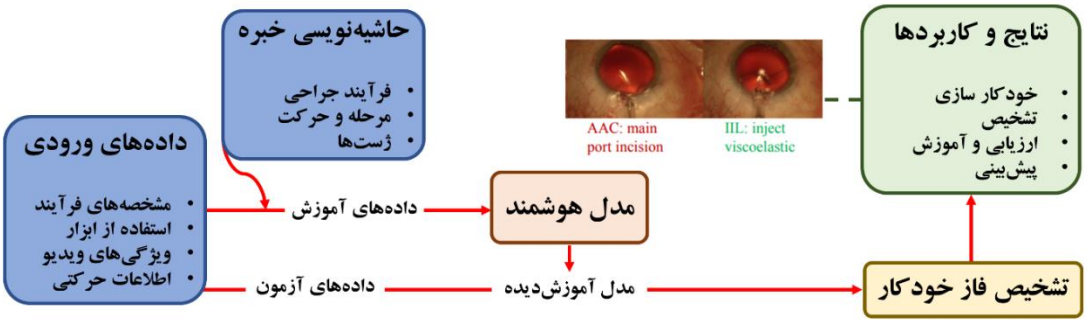
۳- تشخیص مراحل و ژست‌ها از تصاویر جراحی

هر فرآیند جراحی متشکل از چند مرحله و وظیفه است که تشخیص و آگاهی از این مراحل علاوه بر افزودن اطلاعات معنایی برای سیستم‌های خودکار هدایت تصویری جراحی، کمک می‌کند تا پزشکان متخصص در جراحی‌ها یا آموزش آن، از انجام سلسله‌مراتبی و دقیق فرآیندهای جراحی اطمینان یابند. با توسعهٔ ابزارهای رباتیکی کمک‌جراح، تشخیص خودکار فازهای جراحی به ربات کمک خواهد کرد که ربات از مرحله و وظیفهٔ در حال اجرا اطلاعات دقیق‌تری داشته باشد و پاسخ بهتری به اتفاقات زمینه‌ای در حال اجرا یا پیش‌رو ارائه کند. تشخیص فاز جراحی شامل تقسیم یک رویه به مجموعه‌ای از مراحل، و سپس آموزش مدل برای شناسایی فازی از جراحی است که با تصویر داده‌شده مطابقت دارد.

هم‌چنین، با تجزیه و تحلیل حرکات و ژست‌های جراحی به‌جای فازها، می‌توان وظایف جراحی را در سطح عمیق‌تری تجزیه و تحلیل کرد. این فرآیند چالش‌برانگیزتر است؛ چراکه، ژست‌ها در مقایسه با فازها و مراحل کلی، بیش‌تر به یکدیگر شبیه هستند و تشخیص آن‌ها از هم سخت است. فرآیندهای جراحی را می‌توان به مجموعه‌ای از حرکات که ژست‌های مخصوص به خود را دارند تقسیم کرد. مثلاً، فرآیند بخیه‌زنی می‌تواند به مجموعه‌ای از حرکات از عبور سوزن، گره‌زنی و غیره تقسیم شود. هم‌چنین، فرآیند کپسولرکسیس در جراحی آب‌مروارید می‌تواند به حرکاتی مانند برداشتن لایهٔ برگه‌شکل از سطح چشم و حرکت دادن آن در مسیری دایره‌ای‌شکل تقسیم شود. یکی از کارهای دیگری که در این زمینه می‌توان انجام داد، بخش‌بندی مسیر جراحی است. این کار می‌تواند فرآیندهای مختلف دخیل در پردازش تصویر و جراحی، از تشخیص فاز تا ارزیابی مهارت را تسهیل کند. برای بخش‌بندی مسیر معمولاً اطلاعات سینماتیکی استخراج‌شده از ربات‌های جراح را با اطلاعات ویدیویی ادغام می‌کنند. در مجموع این فعالیت کمک می‌کند تا مرحله‌ای که جراح در آن قرار دارد شناسایی شود، حرکات و ژست‌های او مشخص و تحلیل شود، مسیر حرکتی او بخش‌بندی شود و درنهایت بر اساس اطلاعات برآمده از این فعالیت‌ها می‌توان در زمینه‌های کمک به جراح حین جراحی (ارائهٔ بزخورد و راهنمایی مرتبط به فاز جراحی و براساس حرکات و ژست‌های جراح)، ارزیابی مهارت و آموزش جراحی استفاده کرد.

از طرف دیگر، پزشکان متخصص از ضبط ویدیویی روش‌های جراحی برای آموزش جراحان بی‌تجربه یا برای بررسی و یادگیری اشتباهاتی که در طول مداخلات رخ داده است، استفاده می‌کنند. آن‌ها ساعت‌ها از فیلم‌های جراحی را به منظور یافتن صحنه‌های ویدئویی مرتبط با یک فرآیند یا مرحلهٔ جراحی، به‌صورت دستی جستجو و حاشیه‌نویسی می‌کنند. این فرآیند خسته‌کننده و وقت‌گیر است. بنابراین، ضرورت وجود یک ابزار حاشیه‌نویسی (نیمه)خودکار مراحل، حرکات و ژست‌ها در ویدیوهای جراحی احساس می‌شود. توسعهٔ مدل‌هایی برای تشخیص خودکار فازهای جراحی در جهت رفع این نیاز کمک شایانی می‌کند.

از جمله چالش‌های این وظیفه، می‌توان به عدم توافق روی روش‌ها و مراحل کلی جراحی، نیاز به داده‌ی حاشیه‌نویسی‌شدهٔ نسبتاً زیاد و سختی تشخیص دقیق زمان تغییر از یک مرحله به مرحلهٔ دیگر اشاره کرد. نمایی از عمل‌کرد کلی این وظیفه در شکل ۴ نشان داده شده است. در ادامه شماری از کارهای شاخص در این زمینه را معرفی می‌کنیم.



شکل ۴ - مراحل و کاربردهای تشخیص فاز جراحی

پتسچارنیگ و شوفمن [۶۸]، با استفاده از مفهوم یادگیری انتقالی و شبکهٔ کانولوشنی AlexNet پیش‌آموزش‌دیده، مدلی تک‌قابه (تک‌فریمی) برای طبقه‌بندی تصاویر جراحی پیشنهاد کردند که بعدتر، این کار را در سال ۲۰۱۸ [۶۹] توسعه دادند. در مطالعات بعدی [۷۰]، آن‌ها تأثیر ادغام اولیه و دیرهنگام اطلاعات زمانی را در طبقه‌بندی فازهای جراحی بررسی کردند. ادغام اولیه به استخراج اطلاعات حرکتی از دو قاب تصویری ویدیویی متوالی و ادغام آن با تصویر RGB اشاره دارد. با این رویکرد، آن‌ها بیش از ۱۰٪ از کارهای قبلی بهتر عمل کردند.

توینانداً و همکاران [۱۶]، یک معماری جدید CNN به نام EndoNet ارائه کردند که تشخیص فاز و تشخیص حضور ابزار در جراحی را تنها با استفاده از اطلاعات دیداری انجام می‌دهد. EndoNet توسعه‌ای از معماری AlexNet است که در آن آخرین لایه به یک لایهٔ

تماماً متصل منتهی می‌شود که تشخیص ابزار را انجام می‌دهد. سپس خروجی این لایه با خروجی AlexNet الحاق می‌شود تا ویژگی‌های دیداری از تصاویر استخراج شود. سپس، با استفاده از ماشین بردار پشتیبان (SVM) و مدل‌های مارکوف پنهان سلسله مراتبی و این ویژگی‌ها، فاز فعلی جراحی تخمین زده می‌شود. آن‌ها EndoNet را روی مجموعه‌دادهٔ Cholec80 پیاده‌سازی کردند. جین و همکاران [۷۱]، از ارتباط بین تشخیص ابزار و تشخیص فاز با استفاده از یک شبکهٔ یادگیری عمیق چندوظیفه‌ای بهره‌برداری کردند. معماری پیشنهادی آن‌ها دو شاخه دارد. یک ماژول CNN برای تشخیص ابزار و یک ماژول RNN برای تشخیص فاز. آن‌ها برای مدل کردن ارتباط بین این دو کار و به حداقل رساندن واگرایی پیش‌بینی‌های دو شاخه، یک تابع اتلاف مخصوص همبستگی طراحی کردند. آن‌ها در کاری دیگر [۷۲]، یک شبکه با نام SV-RCNet پیشنهاد کردند که اطلاعات دیداری و زمانی را ادغام می‌کند. شبکهٔ پیشنهادی آن‌ها از ResNet-50 برای استخراج ویژگی‌های دیداری و از شبکهٔ LSTM برای مدل‌سازی اطلاعات زمانی موجود در قاب‌های تصویری متوالی استفاده می‌کند. آن‌ها در این کار مزیت یادگیری مشترک ویژگی‌های مکانی-زمانی را در برابر یادگیری جداگانهٔ این ویژگی‌ها نشان دادند.

روش‌های معمول تشخیص خودکار فاز جراحی محدود هستند؛ چراکه، به دلیل ظاهر شدن فریم‌های مشابه در فازهای مختلف ممکن است استخراج ویژگی دیداری با هدف تشخیص فاز به‌صورت مطلوبی انجام نشود. هم‌چنین به دلیل محدودیت‌های محاسباتی که می‌تواند بر تجزیه و تحلیل ویدیوهای طولانی تأثیر بگذارد، ممکن است ویژگی‌های محلی و عام به شکل ضعیفی با هم ترکیب شوند. برای غلبه بر این چالش‌ها، لیو و همکاران [۷۳]، یک ساختار مبتنی بر تبدیل‌کننده‌ها و سازوکار توجه با نام LoViT پیشنهاد کردند که اطلاعات زمانی کوتاه و بلندمدت را با هم در نظر می‌کند و ترکیب می‌کند.

فانک و همکاران [۷۴]، روی‌کردهای مختلف را برای استفاده از مفهوم انسجام زمانی در پیش آموزش شبکه‌های کانولوشنی با هدف بخش‌بندی فازهای جراحی بررسی کردند. آن‌ها فرض کردندکه پیش‌آموزش، CNN را تشویق می‌کند تا ویژگی‌هایی را بیاموزد که در تغییر فاز جراحی معنا دارند و نسبت به تغییرات نامربوط (مانند حرکات جزئی ابزارها یا آندوسکوپ) بین قاب‌های تصویری مجاور تغییر نمی‌کنند. مفهوم انسجام زمانی هم به این معناست که فرض کنیم قاب‌های تصویری به ترتیب گذشت زمان هستند، در طول زمان به آرامی تغییر می‌کنند و تغییرات آن‌ها پیوسته است (تغییرات ناگهانی ندارند).

چیتاجالا و همکاران [۷۵]، هم روشی برای استخراج توصیف‌گرهای محتوای ویدیویی جراحی برای یافتن بخش‌های مشابه و همه در ویدیوهای لاپاراسکوپی اِعمال می‌شود. آن‌ها یک مدل ResNet50 را برای استخراج توصیف‌گرهای تصویری معنایی برای تسهیل جستجو در پایگاه‌های داده بزرگ آموزش دادند. نمازی و همکاران [۷۶]، یک روش یادگیری عمیق برای تشخیص زمان انتقال فازهای مختلف با یادگیری قاب‌های تصویری آغازین و پایانی هر فاز پیشنهاد کردند. کنان و همکاران [۷۷]، چالش طبقه‌بندی ویدیویی برای تشخیص زودهنگام فاز جراحی را با استفاده از معماری CNN-LSTM حل کردند. CNN اطلاعات مکانی را در قاب‌های تصویری ویدیو استخراج و ذخیره می‌کند، در حالی که LSTM اطلاعات زمانی مربوط به تکمیل فرآیند جراحی را مدل و ذخیره می‌کند. از آن‌جایی که هدف این کار تشخیص زودهنگام است، در طول آموزش، CNN با نمونه‌هایی از ویدیوی کامل، با تاکید بر قاب‌های تصویری اولیه با وزن‌های بالاتر تغذیه می‌شود. آن‌ها یک چهارچوب جدید مبتنی بر LSTM برای پیش‌بینی رویدادهای آینده معرفی کردند که عملکرد تشخیص زودهنگام را بهبود بخشیده است.

روش‌های نظارت‌شده به مقدار زیادی دادهٔ حاشیه‌نویسی‌شده نیاز دارند. چن و همکاران [۷۸]، یک روش نیمه‌نظارتی مبتنی بر CNN پیشنهاد کردند. آن‌ها ابتدا از یک شبکهٔ متخاصم مولد (GAN) برای استخراج ویژگی‌های مکانی از تصاویر استفاده کردند. سپس، از یک شبکهٔ LSTM برای تشخیص موضوع و زمینهٔ زمانی قاب‌های تصویری استفاده کردند. در نهایت، آن‌ها از یک روش یادگیری نیمه‌نظارتی برای ادغام اطلاعات مکانی و زمانی برای تنظیم دقیق (تدقیق) شبکه استفاده کردند.

با وجود روش‌های مختلف پیاده‌سازی‌شده در این زمینه، مقایسه معنی دار این روش‌ها به دلیل تفاوت در فرآیند ارزیابی و گزارش ناقص جزئیات ارزیابی دشوار است. به طور خاص، جزئیات محاسبات شاخص می‌تواند به طور گسترده‌ای بین مطالعات مختلف تشخیص فاز جراحی متفاوت باشد. فانک و همکاران [۷۹]، در پژوهشی این چالش را بررسی کردند و راه‌کارهایی برای آن ارائه کردند.

تشخیص حرکات و ژست‌های جراحی فقط بر اساس ویدیو یک مسألهٔ چالش برانگیز است که به ابزارهای موثر برای استخراج اطلاعات دیداری و زمانی از ویدیو نیاز دارد. روی‌کردهای اولیه‌ای که در این زمینه معرفی شدند عمدتاً به استخراج‌کننده‌های ویژگی مبتنی بر قاب تصویری، اعم از دست‌ساز یا آموخته‌شده، متکی هستند که نمی‌توانند پویایی را در ویدیوی جراحی ثبت کنند. فانک و همکاران [۸۰]، برای اولین بار از یک شبکه عصبی کانولوشن سه‌بُعدی (3D-CNN) برای یادگیری ویژگی‌های مکانی-زمانی از قاب‌های تصویری متوالی ویدیو و تشخیص ژست استفاده کردند. خطیبی و دزیانی [۸۱]، یک CNN کم‌عمق را برای تشخیص حرکت جراحی از قاب‌های تصویری تکی و چندتایی پیشنهاد کردند. CNN پیشنهادی آن‌ها را می‌توان با سرعت مناسبی آموزش داد؛ چراکه، نیاز به تنظیم پارامترهای کم‌تری دارد. آن‌ها بهترین عملکرد را با استفاده از قاب‌های تصویری اولیه، میانی و نهایی ویدیوی جراحی به دست آوردند. محدودیت اصلی این کار این است که کارآیی لازم را هنگام استفاده از قاب‌های تصویری‌ای که در آن ابزار جراحی حضور ندارند، ندارد.

لونگو و همکاران [۸۲]، مسألهٔ شناسایی ژست‌ها (تشخیص زمانی که یک ژست اتفاق می‌افتد) و بخش‌بندی ژست‌ها (تشخیص اینکه چه ژستی اتفاق می‌افتد) را در ویدیوهای بخیه‌زنی مورد مطالعه قرار دادند. آن‌ها شبکه‌های مبتنی بر لایه‌های LSTM و CNN-LSTM خود را روی ورودی پیکسل‌های RGB خام و هم‌چنین جریان شار نوری برای هر قاب تصویری آموزش دادند.

نویه و همکاران [۸۳]، یک روش برای تشخیص اقدامات جراحی پیشنهاد کردند که هر فاز جراحی را به سه‌گانهٔ عملِ «ابزار، فعل و هدف» تقسیم می‌کرد. آنها ۴۰ ویدیو از مجموعه‌دادهٔ Cholec80 را با ۶ ابزار، ۸ فعل و ۱۹ کلاس هدف حاشیه‌نویسی کردند و برای تشخیص تعاملات ابزار و بافت از یک شبکهٔ یادگیری عمیق چندکاره با سه شاخهٔ ابزار، فعل و هدف استفاده کردند. شبکهٔ پیشنهادی آن‌ها از لایه‌های کانولوشنی و ساختار ResNet-18 بهره برده است.

مورالی و همکاران [۸۴]، الگوریتمی به نام خوشه‌بندی حالت گذار با یادگیری عمیق (TSC-DL) برای بخش‌بندی وظایف جراحی ارائه کردند. این روش از نوع بدون ناظر است و داده‌ی ویدیویی و سینماتیکی را ادغام می‌کند. آن‌ها از CNNهای پیش‌آموزش‌دیده (AlexNet و VGGNet) برای استخراج ویژگی‌های مرتبط از ویدیوها استفاده می‌کنند و سپس یک فضای حالت افزونه و تقویت‌شده با ویژگی‌های دیداری و داده های سینماتیک ایجاد می‌کنند. نتایج آن‌ها نشان می‌دهد که استفاده از ترکیب اطلاعات سینماتیکی و دیداری عملکرد بهتری نسبت به استفادهٔ صِرف از داده‌ی سینماتیکی دارد. ژائو و همکاران [۸۵]-[۸۶]، هم ساختاری مبتنی بر رمزگذار-رمزگشای کانولوشنی معرفی کرده‌اند که اطلاعات تصویری و سینماتیکی را ادغام می‌کند و از آن‌ها برای کاربرد بخش‌بندی مسیر جراحی استفاده می‌کند. آن‌ها هم‌چنین شاخص‌هایی برای سنجش شباهت بین بخش‌های مسیر پیشنهاد کرده‌اند تا از بخش‌بندی اضافی مسیر جلوگیری کنند.

در جدول ۳ شماری از جدیدترین پژوهش‌های منتشرشده در زمینهٔ تشخیص مراحل، فعالیت‌ها و ژست‌ها در جراحی آورده شده است. همان‌طور که مشخص است، به دلیل این که تشخیص مراحل و ژست‌ها بر اساس تصویر یا توالی تصاویر انجام می‌شود، مدل‌های معروف کانولوشنی در ترکیب با شبکه‌های بازگشتی بسیار مورد استفاده هستند. هم‌چنین در سال‌های اخیر و با معرفی ترنسفورمرها، استفاده از این مدل‌ها رو به گسترش است.

جدول ۳ - شماری از جدیدترین پژوهش‌ها در زمینهٔ تشخیص مراحل، ژست‌ها و فعالیت‌ها در جراحی

هدف	سال و مرجع	مدل بینایی ماشین	مجموعه‌داده	نتایج
تشخیص مرحله	۲۰۲۰ [۷۱]	VGG50 + LSTM	Cholec80 [۱۶]	Accuracy = ۸۹.۲%
تشخیص مرحله	۲۰۲۰ [۲۲]	Xception	LapSig300 [۲۲]	Accuracy = ۸۱%
تشخیص مرحله	۲۰۲۳ [۸۷]	LAST, VAE, Transformer	Cholec80 [۱۶]	Accuracy = ۹۳.۱۲%
تشخیص ژست	۲۰۲۰ [۸۲]	AlexNet + LSTM	اختصاصی	Accuracy = ۸۱%
تشخیص ژست	۲۰۲۳ [۸۸]	Vision Transformer	JIGSAWS [۱۲]	Accuracy = ۸۷.۵%
تشخیص فعالیت	۲۰۲۲ [۸۹]	ASTCFormer, Transformer	اختصاصی (Cholec)	Accuracy = ۹۵.۶۷%

٤- ارزیابی تصویری-ویدیویی مهارت جراحی

با وجود این که تا کنون طیف گسترده‌ای از استانداردهای ارزیابی مهارت جراحی به‌صورت سنتی و ساختاریافته ایجاد شده است و رضایتمندی جراحان استاد و کارآموز تا حدی بهبود یافته است، استفاده از این ساختارها به‌دلیل خودکارنبودن، نیاز به کار فشردهٔ بازبین متخصص، وقت‌گیر و پُرهزینه‌بودن و امکان سوگیری، محدود و غیرفراگیر است [۹۰]. امروزه پزشکان متخصص بر این باورند که علم داده جراحی‌محور باید به عنوان عنصر اصلی در برنامه‌های آموزشی بیمارستان‌هایی که جراحان آینده را آموزش می‌دهند در نظر گرفته شود تا به‌صورت تدریجی رویکردهای هوشمند و مبتنی بر داده، در نسل‌های آینده جایگزین بخش زیادی از زحمات اساتید جراحی گردند. بدین‌ترتیب، اندازه‌گیری مستمر، ارزیابی و بهبود مهارت‌های جراحی با استفاده از ابزار خودکار هدایت تصویری، جزء اصلی نسل بعدی برنامه‌های آموزش جراحی است [۹۱].

ابزارهای ارزیابی سنتی، درک ارزیاب از کیفیت عملکرد جراحی را اندازه گیری می‌کنند؛ درحالی‌که معیارهای خودکار ارزیابی مهارت، عملکرد جراحی را مستقیماً اندازه‌گیری و کمّی می‌کنند. این ابزارها می‌توانند به‌صورت مستقیم از مؤلفه‌های مرتبط به ارزیابی خودکار مهارت جراحی استفاده کنند و یا از مجموعهٔ دانشی برآمده از آن‌ها و الگوریتم‌های هوش مصنوعی و پردازش تصویر برای ارزیابی استفاده کنند. در حالت دوم، این مؤلفه‌ها نقش یک واسطه را برای فهماندن بهتر اطلاعات مرتبط به مهارت جراحی به مدل‌های هوش مصنوعی، ایفا می‌کنند. مطالعات زیادی بررسی کرده‌اند که آیا ادراک ارزیاب می‌تواند با ارزیابی‌های

خودکار انجام‌شده مرتبط باشد یا خیر. نتایج این مطالعات بیان می‌کند که همبستگی بالایی میان این دو شیوهٔ ارزیابی وجود دارد [۶].

تاکنون کارهای مختلفی در زمینهٔ ارزیابی مهارت جراحی با استفاده از داده‌ی حرکتی و نیرویی استخراج‌شده از حسگرهای نصب‌شده روی ابزارهای جراحی انجام شده است. تمرکز این بخش از پژوهش روی کارهایی است که صرفاً از دادهٔ تصویری و روش‌های پردازش تصور و بینایی ماشین استفاده می‌کنند. مزیت این روش‌ها این است که برای بهره‌برداری، نیازی به نصب تجهیزات گران‌قیمت در اتاق‌های عمل ندارند. میکروسکوپ جراحی که در اغلب اتاق‌های عمل موجود است، کار تصویربرداری از جراحی را انجام می‌دهد و صرفاً با تکیه بر همین ابزار و داده می‌توان مهارت جراحان را ارزیابی کرد.

ارزیابی مهارت جراحی با استفاده از روش‌های پردازش تصویر و بینایی ماشین به صورت‌های مختلفی انجام می‌شود. در یکی از روی‌کردهای مرسوم، ویدیوهای جراحی به‌صورت یکپارچه و مستقیم وارد یک شبکهٔ عصبی عمیق می‌شوند تا طبقه‌بندی مهارتی شوند. کاربردهای مربوط به تشخیص فعالیت[1] مشابهت بسیاری با کاربرد تشخیص تصویری-ویدیویی مهارت جراحی دارند. در نوامبر سال ۲۰۱۹، شبکه‌های TSNs[1] توسط ونگ و همکارانش به عنوان یک چهارچوب یادگیری عمیق برای طبقه بندی و تشخیص اعمال انسانی و مبتنی بر ویدیو معرفی شد [۹۲]-[۹۳]. فانک و همکاران [۹۴]، دادهٔ ویدیویی مجموعه‌دادهٔ JIGSAWS را با استفاده از روش‌های یادگیری عمیق و ترکیب یک شبکه ی کانولوشنی سه-بعدی با آموزش TSNمحور طبقه‌بندی مهارتی کردند. ایدهٔ اصلی فانک و همکاران استفاده از شبکه‌های عصبی کانولوشنی برای طبقه‌بندی چندین قطعهٔ کوتاه از یک فیلم، تخصیص امتیاز مهارتی به آن‌ها و سپس پیش‌بینی سطح مهارتی نهایی از طریق اجماع بین نتایج تکه ویدیوها بوده است. در تکه‌ویدیوهای که برای استفاده در شبکهٔ عصبی عمیق در این کار آماده شده اند دو حالت RGB و جریان شار نوری[2] در نظر گرفته شده‌اند. در حالی که کارهای پیشین نشان می‌دهند که استفاده از شار نوری نسبت به RGB به نتایج بهتری در فرآیند تشخیص حرکات انسانی مبتنی بر یادگیری عمیق می انجامد، استفاده از شار نوری هزینهٔ محاسباتی بالایی دارد. هم‌چنین در خروجی شار نوری حرکات اجسام متحرک بین قاب‌های تصویری نمایان شود؛ اما اثری از نواحی ثابت در آن نیست. برای مثال، اگرچه قسمت‌های متحرک در صحنهٔ جراحی مانند حرکات ابزار و بافت‌ها از اهمیت بالایی برخوردارند، برخی بخش‌های صحنه مانند گسترش رنگ قرمز در بافت چشم هم به مهارت جراح در تعامل با بافت ارتباط دارد.

به منظور تسهیل در یادگیری ویژگی‌های فضایی-زمانی[1] در تشخیص و طبقه‌بندی حرکات انسانی، برخی پژوهش‌ها استفاده از شبکه‌های عصبی کانولوشنی سه‌بُعدی[3] را پیشنهاد کرده‌اند [۹۵], [۹۶]. به‌صورت خلاصه شبکه‌های عصبی کانولوشنی سه‌بُعدی نوعی از شبکه‌های کانولوشنی هستند که در آن فیلترهای کانولوشنی، عملگرهایی مانند اجماع و نگاشت‌های ویژگی دارای بُعد سومی به‌صورت زمانی هستند. کریرا و زیسرمن [۹۷]، چگونگی گسترش و تعمیم معماری شبکه‌های کانولوشنی دوبُعدی در امتداد بعد زمانی برای رسیدن به شبکه‌های کانولوشنی سه‌بُعدی را توضیح داده‌اند و یک نسخهٔ سه‌بُعدی از شبکهٔ Inception [۳۱] با نام Inception I3D معرفی کرده‌اند. از آن جا که این شبکه در تشخیص بسیاری از حرکات مختلف انسانی نتیجهٔ امیدوارکننده‌ای داشته است، می‌توان از آن برای کاربرد مهارت حرکات دست در جراحی استفاده کرد. آموزش موفقیت‌آمیز شبکه‌های عصبی عمیق از ابتدا[4] نیازمند به تعداد زیادی داده و منابع محاسباتی است. یک راهبرد محبوب برای حل محدودیت‌های موجود در این زمینه، پیش‌آموزش شبکهٔ عصبی عمیق روی یک مجموعه‌دادهٔ بزرگ برچسب‌گذاری‌شده از دامنهٔ مرتبط با حوزهٔ مقصد (ارزیابی مهارت جراحی) و سپس تنظیم دقیق[5] پارامترها پس از این مرحله است [۹۷]. فانک و همکاران از یک شبکهٔ Inception I3D پیش‌آموزش‌دیده روی مجموعه‌دادهٔ کینتیک[6] [۹۸] که به‌صورت عمومی برای هر دو شکل ورودی RGB و جریان شار نوری در دسترس است، استفاده کرده‌اند. این مجموعه‌داده شامل ۴۰۰ کلاس از کنش‌های انسانی است که در هر یک از این کلاس‌ها حداقل ۴۰۰ ویدیو وجود دارد. شبکهٔ عصبی عمیق با پیش‌آموزش روی این مجموعه‌داده، بسیاری از ویژگی‌های فضایی-زمانی نهفته در ویدیوها که به حرکات انسانی مربوط است را فراگرفته است. با شروع از آن نقطه به‌جای شروع تصادفی اوزان و سایر پارامترها، می‌توان با هزینهٔ محاسباتی و داده کم‌تری در حوزهٔ مقصد به نتیجه‌ای مطلوب دست پیدا کرد. فانک و همکاران برا کاهش پارامترهای قابل آموزش و سبک‌سازی فرآیند آموزش، اوزان تمامی لایه‌ها در فرآیند تنظیم دقیق را به جز لایهٔ دوم تا آخر بلوک Mixed-5b ثابت[1] نگه داشتند. با انجام این کارها فانک و همکاران توانستند به دقت ۹۵ درصد در طبقه‌بندی مهارتی داده مجموعه‌دادهٔ JIGSAWS در سه دستهٔ تازه‌کار، متوسط و ماهر دست پیدا کنند.

کیتاگوچی و همکاران [۹۹]، بر اساس ساختار پیشنهادی فانک و همکاران، یک پیاده‌سازی روی مجموعه‌داده‌ای واقعی از جراحی انجام دادند. داده‌ی ویدیویی استفاده‌شده در این کار شامل ویدیوهایی از جراحی لاپاراسکوپی رودهٔ بزرگ بوده است. نتایج مربوط به ارزیابی مهارت ویدیویی این مجموعه‌داده در وظایف مختلف بین ۷۵ تا ۹۰ درصد گزارش شده است. این نشان می‌دهد که ساختار معرفی‌شده در پژوهش فانک و همکاران قدرت تعمیم‌پذیری امیدبخشی دارد؛ اما در هر صورت عمل‌کرد بدتری در ویدیوهای واقعی جراحی نسبت به مجموعه‌دادهٔ شبیه‌سازی‌شدهٔ JIGSAWS دارد. داتلی و همکاران [۱۰۰] هم ساختاری مبتنی بر TSN پیشنهاد کردند که قابلیت استنباط تفاوت‌های مهارتی در ویدیوهای

[1] Action Recognition
[2] Optical Flow
[3] 3D ConvNets
[4] From Scratch
[5] Fine-tune
[6] Kinetics Human Action Video Dataset

جراحی را دارد. آن‌ها برای هدف خود تابع اتلافی مخصوص هم پیشنهاد کردند.

لیو و همکاران [۱۰۱]، ضمن اشاره به سختی و چالش‌های دست‌یابی به داده‌ی حرکتی جراحی، استفاده از جایگزین‌های ساده مانند داده‌ی دیداری و شاخص وضوح میدان عمل[1] را به‌عنوان روشی مناسب و کم‌هزینه برای ارزیابی مهارت جراحان پیشنهاد کردند. آن‌ها این شاخص را بر مبنای میزان خون‌ریزی و قابل مشاهده بودن اندام‌ها و بافت‌ها تعریف کردند. آن‌ها بیان کردند که یک جراح باتجربه، خون‌ریزی را به حداقل می‌رساند و میدان دید دوربین را شفاف و واضح حفظ می‌کند. سپس با بررسی‌های پزشکی نشان دادند که این شاخص ارتباط بالایی با میزان مهارت و تجربهٔ جراحان دارد. آن‌ها با استفاده از روش‌های مبتنی بر پردازش تصویر، ویژگی‌های رنگی و معنایی ویدیوهای جراحی لاپاراسکوپی را استخراج کرده و آن‌ها را به شبکه‌های عصبی وارد کردند تا امتیازات لازم برای هر ویدیو محاسبه شود.

بسیاری از پژوهش‌های انجام‌گرفته در زمینهٔ ارزیابی مهارت ویدیویی، از شبکه‌های کانولوشنی و بازگشتی برای استخراج ویژگی‌ها و روابط کوتاه و بلندمدت استفاده کرده‌اند. برای مثال، ونگ و همکاران [۱۰۲]، ویژگی‌های دوبُعدی یا سه‌بُعدی را از قاب‌های تصویری ویدیو و با استفاده از شبکه‌های کانولوشنی استخراج کردند و با استفاده از شبکهٔ بازگشتی، رابطهٔ زمانی آن‌ها را مدل‌سازی کردند و از آن برای طبقه‌بندی مهارت و ژست‌های جراحی استفاده کردند.

بیش‌ترِ روش‌های پیاده‌سازی‌شده در این زمینه، قبل از آن‌که ویژگی‌ها را به شبکه یا ساختار بعدی وارد کنند، آن‌ها را در بُعد فضایی (محلی) ادغام می‌کنند. این کار، واریانس‌های معنایی ویژگی‌های مختلف را نادیده می‌گیرد و تمام اطلاعات و ویژگی‌ها را در مقیاس فضایی بدون تمایز با هم فشرده می‌سازد. در نتیجه، شبکه‌ها و ساختارهای متعاقب به سختی می‌توانند رابطهٔ زمانی ویژگی‌های محلی در بخش‌های فضایی مختلف را به‌طور جداگانه مدل‌سازی کنند. به عنوان مثال، حرکات ابزارهای مختلف و تغییرات وضعیت بافت هرچند در ویژگی‌های ادغام‌شده لحاظ می‌شوند؛ اما بروز مشخصی ندارند. این چالش به‌ویژه برای ارزیابی مهارت‌های جراحی مهم است؛ چراکه، ردیابی ابزارها در بخش بزرگی از پس‌زمینه، برای قضاوت در مورد کیفیت جراحی ضروری است. به طور مشابه، تعامل بین ابزار و بافت در طول زمان نیز برای ارزیابی مهم است. بسیاری از روش‌های موجود شبکه‌های عصبی سرتاسری هستند که اطلاعات کمی از حرکت یا ظاهری که گرفته می‌شود را آشکار می‌کنند.

برای حلّ این چالش، از آن‌جایی که ویدیوهای جراحی شامل اشیاء محدودی با معانی معنایی صریح مانند ابزار، بافت و پس‌زمینه هستند، ژنکیانگ و همکاران [۱۰۳] یک چهارچوب جدید با نام تجمیع معنایی ویدیویی (ViSA)[2] پیشنهاد کردند که با جمع‌آوری ویژگی‌های محلی در ابعاد مکانی-زمانی و بر اساس سرنخ‌های معنایی، فرآیند ارزیابی مهارت جراحی را انجام می‌دهد. با توجه‌کردن به معنای ویژگی‌ها در طول زمان و مطابقت معناشناسانهٔ بین بخش‌های مختلف ویدیو، شفافیت شبکه بهبود می‌یابد. این ساختار پیشنهادی از شبکه‌های LSTM دوطرفه استفاده کرده است و روی مجموعه‌دادهٔ JIGSAWS به نتایج امیدبخشی دست پیدا کرده است.

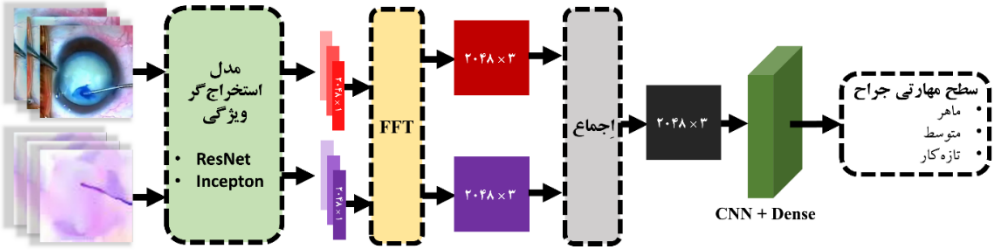
در روشی دیگر برای ارزیابی تصویری مهارت جراحی، ویدیوهای جراحی به قاب‌های تصویری تبدیل می‌شوند. سپس از این قاب‌های تصویری به‌صورت‌های مختلفی نمونه‌برداری می‌شود تا به‌عنوان ورودی یک شبکهٔ عصبی عمیق استفاده شوند. در این مرحله می‌توان از برخی ایده‌ها مانند تبدیل ویدیو به قاب‌های تصویری تفریقی استفاده کرد. قاب تصویری تفریقی حاصل تفریق دو قاب تصویری متوالی است. از آن‌جا که در ارزیابی مهارت، تغییرات روند حرکتی در ویدیو اهمیت دارد، این روش می‌تواند به برجسته‌سازی نکات پراهمیت ویدیو و کاهش هزینهٔ محاسباتی کمک کند. هم‌چنین میزان نمونه‌برداری از قاب‌های تصویری در یک ویدیو را می‌توان با توجه به محدودیت‌های سخت‌افزاری تنظیم کرد. در ادامه، قاب‌های تصویریِ انتخابی به یک شبکهٔ عصبی کانولوشنی عمیق وارد می‌شوند تا یک بردار ویژگی از هر قاب تصویری استخراج شود. این شبکهٔ کانولوشنی می‌تواند پشته‌ای از لایه‌های دلخواه باشد و یا یک شبکهٔ پیش‌آموزش‌دیده که قابلیت استخراج ویژگی بهتری دارد. در گام بعدی، بردارهای ویژگی استخراج‌شده از قاب‌های تصویری هر ویدیو در کنار یکدیگر قرار می‌گیرند و یک ماتریس ویژگی را می‌سازند. حال می‌توان این ماتریس ویژگی را که در هر سطر خود روند تغییرات یک ویژگی استخراج‌شده در طول زمان و قاب‌های تصویری را دارد به یک شبکهٔ جدید وارد کرد و در خروجی آن با قرار دادن لایه‌های تماماً‌متصل ماتریس‌های ویژگی و ویدیوهای متناظر با آن‌ها را در دسته‌های مختلف مهارتی طبقه‌بندی کرد. سلیمانی و همکاران [۱۰۴]، با در پیش گرفتن روی‌کردی مشابه به ارزیابی ویدیویی مهارت جراحی روی مجموعه‌دادهٔ JIGSAWS پرداختند و به دقت ۹۷ درصد دست یافتند.

با پیاده‌سازی این روش برای مجموعه‌دادهٔ JIGSAWS، Cataract-101 و ARAS-Farabi می‌توان به F1 score بین ۶۵ الی ۸۰ درصد دست پیدا کرد. برای بهبود نتایج می‌توان ویژگی‌های استخراج‌شده را از فیلترهای مختلفی مانند فیلتر فرکانسی تبدیل فوریهٔ سریع[3] عبور داد. هم‌چنین در ورودی می‌توان علاوه بر قاب‌های تصویری استخراج‌شده از ویدیوی عادی RGB از قاب‌های تصویری تفریقی و یا قاب‌های تصویری استخراج‌شده از خروجیِ جریان شار نوری ویدیوها استفاده کرد.

نمایی کلی از این فرآیند در شکل ۵ نشان داده شده است. در این شکل ویدیوها به دو صورت RGB و جریان شار نوری به شبکه تغذیه شده اند. در پیاده‌سازی انجام‌شده مشخص شد که با استفاده از این روش‌های می‌توان امتیاز F1 را تا ۹۵ درصد هم ارتقا داد. لازم به ذکر است

[1] Clearness of Operating Field (COF)
[2] Video Semantic Aggregation (ViSA)
[3] Fast Fourier Transform (FFT)

که با تغییر شبکهٔ مبنای استخراج ویژگی کیفیت نتایج هم تغییر خواهد کرد. با پیاده‌سازی‌ها و بررسی‌های انجام‌شده در این پژوهش مشخص شد که شبکه‌های پیش‌آموزش‌دیدهٔ ResNet50 [۱۰۵]، Inception [۳۱] و Xception [۱۰۶] توانایی مناسبی برای استخراج ویژگی مورد نیاز جهت درک نکات کلیدی مهارتی موجود در ویدیوهای جراحی دارند.



شکل ۵ - نمایی از یکی از روش‌های ارزیابی ویدیویی مهارت جراحی

روی‌کرد دیگری که از آن می‌توان برای ارزیابی تصویری-ویدیویی جراحی استفاده کرد، استخراج اطلاعات حرکتی از ویدیوی جراحی و سپس ارزیابی مهارت جراحی بر مبنای این اطلاعات است. برای نمونه، در جراحی کپسولرکسیس با ردیابی سوزن جراحی و قرنیهٔ چشم می‌توان اطلاعات حرکتی آن‌ها را به‌صورت پیکسلی و مطلق به دست آورد. با در دست داشتن این اطلاعات می‌توان وضعیت حرکتی نوک سوزن نسبت به مرکز قرنیه را هم استخراج کرد. سرعت و شتاب هم به سادگی از روی اطلاعات مکانی و زمانی قابل استخراج هستند. از آن‌جا که در ارزیابی مهارت جراحی و ارائهٔ بازخورد به جراحان این روند اطلاعات حرکتی است که اهمیت دارد، می‌توان از این اطلاعات برای تحلیل جراحی استفاده کرد. گو [۱۰۷] و کیم [۱۰۸]، از روی‌کرد مشابهی برای ایجاد شاخص‌های کمّی ارزیابی مهارت در جراحی با استفاده از تحلیل حرکات برمبنای تصویر و ویدیو استفاده کردند.

جراحی که از سطح مهارتی پایین‌تری برخوردار است، لرزش‌ها و حرکات نامطئن بیش‌تری از خود نشان می‌دهد. این نوع از حرکات در فضای زمانی خود را با تغییرات متعدد در علامت سرعت و یا آنتروپی حرکتی نشان می‌دهند که با استفاده از روش‌های ردیابی و پردازش تصویر، به‌دقت از اطلاعات استخراج‌شده از ویدیو قابل محاسبه هستند. هم‌چنین در فضای فرکانسی هم این نوع از حرکات خود را با ضرایب فرکانس‌بالا نشان می‌دهند.

در گروه رباتیک ارس مدل‌های یادگیری عمیق و نرم‌افزاری برای استخراج اطلاعات حرکتی از ویدیوی جراحی و ارزیابی مهارت جراحان طراحی شده است. نرم‌افزار طراحی‌شده بسیاری از چالش‌هایی که در ویدیوی جراحی مانع از تحلیل درست می‌شوند را با استفاده از روش‌های پردازش تصویر حل کرده است. مثلاً انعکاس نور در یک جراحی چشم که شباهت زیادی با نوک سوزن جراحی پیدا می‌کند، با استفاده روش‌های پردازش تصویر مبتنی بر ردیابی حذف می‌شوند. هم‌چنین نرم‌افزار این قابلیت را دارد که ضمن استخراج اطلاعات حرکتی و پیاده‌سازی تبدیلات مختلف، مهارت جراحان را در شاخص‌های مختلفی به‌صورت دقیق و کمّی محاسبه کند. اطلاعات بیش‌تر مربوط به این نرم‌افزار از طریق این پیوند[1] در دسترس است.

۵- جمع‌بندی و نگاه به آینده

علی‌رغم پیچیدگی‌های فرآیندهای جراحی، به لطف پیشرفت‌های روش‌های هوش مصنوعی، پردازش تصویر و بینایی ماشین در سال‌های گذشته، پیشرفت قابل‌توجهی در سیستم‌های خودکار هدایت تصویری جراحی پدید آمده است. جامعهٔ علمی در زمینهٔ استخراج اطلاعات معنایی از تصاویر جراحی به منظور ارائه آگاهی از زمینه در سناریوهای پیچیدهٔ واقعی، پیشرفت های عظیمی داشته است. با این حال، هنوز کارهای زیادی باید انجام شود تا سیستم‌های رباتیکی و رایانه‌ای درکی واقعی و کارآمد از صحنهٔ جراحی داشته باشند.

ایده‌های بسیاری هم‌چون ایدهٔ اروپایی Smart Autonomous Robotic Assistant Surgeon [۱۰۹] مطرح هستند که هدف آن‌ها توسعهٔ یک سیستم رباتیکی است که از طریق سامانه‌های مبتنی بر پردازش تصویر و بینایی ماشین قادر به درک عمیق سناریوی جراحی و همکاری مستقل با جراح باشد. برای دستیابی به این اهداف، تشخیص و بخش‌بندی خودکار مراحل، ابزارها، بافت‌ها و ژست‌ها ضروری است. هم‌چنین، توسعهٔ ابزارهای ارزیابی ویدیویی مهارت جراحی کمک می‌کنند که این سامانه‌ها یک فهم سطح بالا از آن‌چه در یک جراحی مطلوب است و آن‌چه در یک جراحی مطلوب نیست داشته باشند و از طریق آن یک دانش اندوخته برای همکاری یا انتقال مهارت در جراحی کسب کنند.

بر اساس بررسی ارائه‌شده در این پژوهش، می توان به این جهت تحقیقاتی امیدوارکننده اشاره نمود: توسعه سامانه‌های کمک جراحی و سامانه‌های آموزش جراحی هوشمند، که هر دو به درک سناریوی جراحی و ابزار هدایت تصویری نیازی ضروری خواهند داشت. خطوط تحقیقاتی

[1] aras.kntu.ac.ir/softwares

مطرح‌شده در این پژوهش، نقشی کلیدی در افزایش قابلیت‌های روش‌های جراحی فعلی خواهد داشت، که به نفع پرسنل پزشکی و بیماران است.

زمینهٔ تشخیص، ردیابی و بخش‌بندی در جراحی، با معرفی ساختارهایی مانند SAM که حتی بدون آموزش قادر به بخش‌بندی در هر صحنه‌ای هستند، دچار تحول بسیاری خواهد شد. با توسعهٔ مدل‌هایی بر این مبنا و افزایش کارآیی آن‌ها برای صحنه‌های جراحی با استفاده از روش‌هایی مانند تدقیق، ابزارها و بافت‌ها در ویدیوهای جراحی‌های مختلف امکان بخش‌بندی با دقت بالایی خواهند داشت. با این توسعه، پیش‌بینی می‌شود که سامانه‌های رباتیکی و سامانه‌های هدایت تصویری جراحی به‌صورتی فراگیر مجهز به ابزارهای بینایی ماشین لازم برای تشخیص، بخش‌بندی و ردیابی شوند و از طریق این ابزارها و بدون نیاز به نصب حس‌گرهای گران‌قیمت، با دقت مناسبی به استخراج اطلاعات حرکتی از صحنهٔ جراحی بپردازند. هم‌چنین با توجه به توسعهٔ این مدل‌های قدرتمند می‌توان ادعا کرد که اکنون با جمع‌آوری داده‌های مناسب از یک جراحی، این امکان وجود خواهد داشت تا با ایجاد تغییرات جزئی و تنظیم مدل‌های موجود به دقت‌های مناسبی دست پیدا کرد. از جمله چالش‌هایی که این زمینه هم‌چنان با آن روبرو خواهد بود می‌توان به تشخیص، بخش‌بندی و ردیابی بسیار دقیق یک نقطهٔ مشخص از بافت یا ابزار جراحی مانند نوک سوزن جراحی اشاره کرد. با ترکیب روی‌کردهای کلاسیک پردازش تصویر، ریاضیات و جدیدترین مدل‌های بینایی ماشین می‌توان بر این چالش‌ها غلبه کرد.

زمینهٔ تشخیص مراحل، فعالیت‌ها و ژست‌های جراحی هم با جمع‌آوری هزاران تصویر از صحنه‌های جراحی و با توسعهٔ ساختارهایی مانند مدل‌های مبتنی بر ترنسفورمر به زودی به یک قسمت مهم و جدانشدنی از سامانه‌های هدایت تصویری جراحی تبدیل خواهد شد. این سامانه‌ها با استفاده از این زمینه و اطلاعات حرکتی استخراج‌شده از ویدیو قادر خواهند بود که به فهمی مناسب از آن‌چه در یک صحنهٔ جراحی در جریان است دست پیدا کنند. در این زمینه، تشخیص ژست در ترکیب با مدل‌هایی که قادر به درک مهارت جراحی هستند می‌تواند به یک ابزار برای تشخیص و انتقال سبک یک جراح متخصص به کارآموزانش تبدیل شود.

پیشرفت در زمینهٔ ارزیابی تصویری-ویدیویی مهارت جراحی نسبت به زمینه‌های دیگر با سختی بیش‌تری روبه‌رو است. تاکنون، کارهای مختلفی در زمینهٔ تشخیص حرکات و فعالیت‌های انسانی انجام شده است؛ اما ارزیابی مهارت جراحی از روی ویدیو نیازمند یک فهم بسیار عمیق از صحنهٔ جراحی است. تمامی ویدیوهای یک جراحی مشخص دارای صحنه‌ها و ویژگی‌های دیداری تقریباً مشابه هستند. بنابراین، تشخیص تفاوت‌های مهارتی در این صحنه‌های از نظر دیداری تقریباً مشابه کاری چالش‌برانگیز است. پروژه‌های مختلفی در دانشگاه‌ها، بیمارستان‌ها و شرکت‌های دولتی و خصوصی دنیا در این زمینه در حال انجام است. با به ثمررسیدن این تلاش‌ها و طراحی مدلی که قادر به فهم مهارتی در ارزیابی مهارت جراحان باشد، تمام رفتارهایی که مربوط به یک جراحی ماهرانه است و تمام رفتارهایی که در مقابل رفتار ماهرانه قرار دارد در یک مدل هوشمند غیرانسانی نهفته‌سازی خواهد شد. این مدل، علاوه بر ارزیابی مهارت جراحی، به سامانه‌های هدایت تصویری جراحی برای تشخیص و پیش‌بینی در یک صحنهٔ جراحی و مداخلهٔ موثر در آن کمک بسیار خواهد کرد. هم‌چنین می‌توان از این مدل‌های مهارت‌فهم، برای انتقال مهارت یا یک سبک خاص از جراحی جراحی به کارآموزان جراحی استفاده کرد و فرآیند آموزش جراحی را تا حد بسیار زیادی به‌صورت خودکار درآورد.

تمامی اجزای معرفی‌شده در این پژوهش، در کنار یکدیگر منجر به ایجاد درکی عمیق از صحنهٔ جراحی خواهند شد. درک صحنهٔ جراحی برای همکاری انسان در طول یک جراحی واقعی و تصمیم‌گیری در مورد شیوه و زمانی که سامانه باید اقدام خاصی را انجام دهد ضروری است. ضمن اهمیت توسعهٔ هریک از این اجزا به‌صورت مجزا، بهم‌پیوستن این اجزا در یک قالب یکپارچه نیز از اهمیت بالایی برخورد است. در این زمینه، استفاده از مدل‌های هوش مصنوعی سلسله‌مراتبی و برخی شبکه‌ها، مانند شبکه‌های عصبی گراف کمک‌کننده خواهد بود.

علاوه بر این موضوعات، پس از همه‌گیری بیماری کرونا در سال ۲۰۲۰ میلادی، استفاده از روی‌کردهای مبتنی بر واقعیت مجازی و واقعیت افزوده، و هم‌چنین خودکارسازی بسیاری از مراحل جراحی و آموزش آن، از اهمیت بسیار بالاتری برخوردار شده است. با توسعهٔ استفاده از پردازش تصویر و بینایی ماشین در جراحی، تمام بسترهای لازم برای این مهم عملیاتی خواهد شد. پژوهش‌های بررسی‌شده در این مقاله و شماری پژوهش دیگر با تقسیم‌بندی بیش‌تر و به‌همراه کدهای پیاده‌سازی در پیوند گیت‌هاب[1] در دسترس هستند.

تصدیق

بخشی از هزینه های این پژوهش توسط کمک‌هزینه (گرنت) مشترک NIMAD-INSF به شمارهٔ ۴۰۰۱۱۹۰ (NIMAD) و ۴۰۰۲۷۶۶ (INSF) مورد پشتیبانی قرار گرفته است.

مراجع

[1] A. Gawande, "Two Hundred Years of Surgery," New England Journal of Medicine, vol. 366, no. 18, 2012, doi: 10.1056/nejmra1202392.

[2] M. Swaminathan, S. Ramasubramanian, R. Pilling, J. Li, and K. Golnik, "ICO-OSCAR for pediatric cataract surgical skill assessment," Journal of AAPOS, vol. 20, no. 4, 2016, doi: 10.1016/j.jaapos.2016.02.015.

[3] L. Maier-Hein et al., "Surgical data science – from concepts toward clinical translation," Medical Image Analysis, vol. 76. 2022. doi: 10.1016/j.media.2021.102306.

[1] github.com/MJAHMADEE/Awesome_Video-guided_Surgery

A Feasibility Study," IEEE Robot Autom Lett, vol. 5, no. 4, 2020, doi: 10.1109/LRA.2020.3013914.

[16] A. P. Twinanda, S. Shehata, D. Mutter, J. Marescaux, M. De Mathelin, and N. Padoy, "EndoNet: A Deep Architecture for Recognition Tasks on Laparoscopic Videos," IEEE Trans Med Imaging, vol. 36, no. 1, 2017, doi: 10.1109/TMI.2016.2593957.

[17] I. Aksamentov, A. P. Twinanda, D. Mutter, J. Marescaux, and N. Padoy, "Deep neural networks predict remaining surgery duration from cholecystectomy videos," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2017. doi: 10.1007/978-3-319-66185-8_66.

[18] A. Jin et al., "Tool detection and operative skill assessment in surgical videos using region-based convolutional neural networks," in Proceedings - 2018 IEEE Winter Conference on Applications of Computer Vision, WACV 2018, 2018. doi: 10.1109/WACV.2018.00081.

[19] K. Schoeffmann, H. Husslein, S. Kletz, S. Petscharnig, B. Muenzer, and C. Beecks, "Video retrieval in laparoscopic video recordings with dynamic content descriptors," Multimed Tools Appl, vol. 77, no. 13, 2018, doi: 10.1007/s11042-017-5252-2.

[20] A. Leibetseder et al., "LapGyn4: A dataset for 4 automatic content analysis problems in the domain of laparoscopic gynecology," in Proceedings of the 9th ACM Multimedia Systems Conference, MMSys 2018, 2018. doi: 10.1145/3204949.3208127.

[21] A. Leibetseder, S. Kletz, K. Schoeffmann, S. Keckstein, and J. Keckstein, "GLENDA: Gynecologic Laparoscopy Endometriosis Dataset," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2020. doi: 10.1007/978-3-030-37734-2_36.

[22] D. Kitaguchi et al., "Automated laparoscopic colorectal surgery workflow recognition using artificial intelligence: Experimental research," International Journal of Surgery, vol. 79, 2020, doi: 10.1016/j.ijsu.2020.05.015.

[23] M. J. Primus et al., "Frame-Based Classification of Operation Phases in Cataract Surgery Videos," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2018. doi: 10.1007/978-3-319-73603-7_20.

[24] K. Schoeffmann, M. Taschwer, S. Sarny, B. Münzer, M. J. Primus, and D. Putzgruber, "Cataract-101 - Video dataset of 101 cataract surgeries," in Proceedings of the 9th ACM Multimedia Systems Conference, MMSys 2018, 2018. doi: 10.1145/3204949.3208137.

[4] T. G. Weiser et al., "An estimation of the global volume of surgery: a modelling strategy based on available data," The Lancet, vol. 372, no. 9633, 2008, doi: 10.1016/S0140-6736(08)60878-8.

[5] T. G. Weiser et al., "Estimate of the global volume of surgery in 2012: an assessment supporting improved health outcomes," The Lancet, vol. 385, 2015, doi: 10.1016/s0140-6736(15)60806-6.

[6] N. Alnafisee, S. Zafar, S. S. Vedula, and S. Sikder, "Current methods for assessing technical skill in cataract surgery," Journal of Cataract and Refractive Surgery, vol. 47, no. 2. 2021. doi: 10.1097/j.jcrs.0000000000000322.

[7] M. Nathan et al., "Intraoperative adverse events can be compensated by technical performance in neonates and infants after cardiac surgery: A prospective study," Journal of Thoracic and Cardiovascular Surgery, vol. 142, no. 5, 2011, doi: 10.1016/j.jtcvs.2011.07.003.

[8] S. E. Regenbogen, C. C. Greenberg, D. M. Studdert, S. R. Lipsitz, M. J. Zinner, and A. A. Gawande, "Patterns of Technical Error Among Surgical Malpractice Claims," Ann Surg, vol. 246, no. 5, 2007, doi: 10.1097/sla.0b013e31815865f8.

[9] R. Anteby et al., "Deep learning visual analysis in laparoscopic surgery: a systematic review and diagnostic test accuracy meta-analysis," Surgical Endoscopy, vol. 35, no. 4. 2021. doi: 10.1007/s00464-020-08168-1.

[10] K. Nandigam, J. Soh, W. G. Gensheimer, A. Ghazi, and Y. M. Khalifa, "Cost analysis of objective resident cataract surgery assessments," J Cataract Refract Surg, vol. 41, no. 5, 2015, doi: 10.1016/j.jcrs.2014.08.041.

[11] Anaconda, "The State of Data Science 2020 Moving from hype toward maturity, 2020."

[12] Y. Gao et al., "JHU-ISI Gesture and Skill Assessment Working Set (JIGSAWS): A Surgical Activity Dataset for Human Motion Modeling," Modeling and Monitoring of Computer Assisted Interventions (M2CAI) – MICCAI Workshop, 2014.

[13] D. Sarikaya, J. J. Corso, and K. A. Guru, "Detection and Localization of Robotic Tools in Robot-Assisted Surgery Videos Using Deep Neural Networks for Region Proposal and Detection," IEEE Trans Med Imaging, vol. 36, no. 7, 2017, doi: 10.1109/TMI.2017.2665671.

[14] E. Colleoni, P. Edwards, and D. Stoyanov, "Synthetic and Real Inputs for Tool Segmentation in Robotic Surgery," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2020. doi: 10.1007/978-3-030-59716-0_67.

[15] A. Attanasio et al., "Autonomous Tissue Retraction in Robotic Assisted Minimally Invasive Surgery -

[25] M. Grammatikopoulou et al., "CaDIS: Cataract dataset for surgical RGB-image segmentation," Med Image Anal, vol. 71, 2021, doi: 10.1016/j.media.2021.102053.

[26] M. J. Ahmadi et al., "ARAS-Farabi Experimental Framework for Skill Assessment in Capsulorhexis Surgery," in 9th RSI International Conference on Robotics and Mechatronics, ICRoM 2021, 2021. doi: 10.1109/ICRoM54204.2021.9663494.

[27] M. Lafouti et al., "Surgical Instrument Tracking for Capsulorhexis Eye Surgery Based on Siamese Networks," in 10th RSI International Conference on Robotics and Mechatronics, ICRoM 2022, 2022. doi: 10.1109/ICRoM57054.2022.10025355.

[28] D. Bouget, R. Benenson, M. Omran, L. Riffaud, B. Schiele, and P. Jannin, "Detecting Surgical Tools by Modelling Local Appearance and Global Shape," IEEE Trans Med Imaging, vol. 34, no. 12, 2015, doi: 10.1109/TMI.2015.2450831.

[29] S. Wang, A. Raju, and J. Huang, "Deep learning based multi-label classification for surgical tool presence detection in laparoscopic videos," in Proceedings - International Symposium on Biomedical Imaging, 2017. doi: 10.1109/ISBI.2017.7950597.

[30] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings, 2015.

[31]C. Szegedy et al., "Going deeper with convolutions," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2015. doi: 10.1109/CVPR.2015.7298594.

[32] K. Mishra, R. Sathish, and D. Sheet, "Learning Latent Temporal Connectionism of Deep Residual Visual Abstractions for Identifying Surgical Tools in Laparoscopy Procedures," in IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 2017. doi: 10.1109/CVPRW.2017.277.

[33] H. Al Hajj, M. Lamard, P. H. Conze, B. Cochener, and G. Quellec, "Monitoring tool usage in surgery videos using boosted convolutional and recurrent neural networks," Med Image Anal, vol. 47, 2018, doi: 10.1016/j.media.2018.05.001.

[34] X. Du et al., "Articulated multi-instrument 2-d pose estimation using fully convolutional networks," IEEE Trans Med Imaging, vol. 37, no. 5, 2018, doi: 10.1109/TMI.2017.2787672.

[35] E. Colleoni, S. Moccia, X. Du, E. De Momi, and D. Stoyanov, "Deep Learning Based Robotic Tool Detection and Articulation Estimation with Spatio-Temporal Layers," IEEE Robot Autom Lett, vol. 4, no. 3, 2019, doi: 10.1109/LRA.2019.2917163.

[36] Z. Chen, Z. Zhao, and X. Cheng, "Surgical instruments tracking based on deep learning with lines detection and spatio-temporal context," in Proceedings - 2017 Chinese Automation Congress, CAC 2017, 2017. doi: 10.1109/CAC.2017.8243236.

[37] Z. Zhao, T. Cai, F. Chang, and X. Cheng, "Real-time surgical instrument detection in robot-assisted surgery using a convolutional neural network cascade," in Healthcare Technology Letters, 2019. doi: 10.1049/htl.2019.0064.

[38] Y. Liu, Z. Zhao, F. Chang, and S. Hu, "An anchor-free convolutional neural network for real-time surgical tool detection in robot-assisted surgery," IEEE Access, vol. 8, 2020, doi: 10.1109/ACCESS.2020.2989807.

[39] B. Ran, B. Huang, S. Liang, and Y. Hou, "Surgical Instrument Detection Algorithm Based on Improved YOLOv7x," Sensors, vol. 23, no. 11, 2023, doi: 10.3390/s23115037.

[40] C. I. Nwoye, D. Mutter, J. Marescaux, and N. Padoy, "Weakly supervised convolutional LSTM approach for tool tracking in laparoscopic videos," Int J Comput Assist Radiol Surg, vol. 14, no. 6, 2019, doi: 10.1007/s11548-019-01958-6.

[41] M. Islam, Y. Li, and H. Ren, "Learning Where to Look While Tracking Instruments in Robot-Assisted Surgery," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2019. doi: 10.1007/978-3-030-32254-0_46.

[42] X. Du et al., "Patch-based adaptive weighting with segmentation and scale (PAWSS) for visual tracking in surgical video," Med Image Anal, vol. 57, 2019, doi: 10.1016/j.media.2019.07.002.

[43] I. Laina et al., "Concurrent segmentation and localization for tracking of surgical instruments," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2017. doi: 10.1007/978-3-319-66185-8_75.

[44] L. C. Garcia-Peraza-Herrera et al., "ToolNet: Holistically-nested real-time segmentation of robotic surgical tools," in IEEE International Conference on Intelligent Robots and Systems, 2017. doi: 10.1109/IROS.2017.8206462.

[45] A. A. Shvets, A. Rakhlin, A. A. Kalinin, and V. I. Iglovikov, "Automatic Instrument Segmentation in Robot-Assisted Surgery using Deep Learning," in Proceedings - 17th IEEE International Conference on Machine Learning and Applications, ICMLA 2018, 2019. doi: 10.1109/ICMLA.2018.00100.

[46] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in

vol. 6, no. 2, 2021, doi: 10.1109/LRA.2021.3066834.

[57] J. Choi, S. Cho, J. W. Chung, and N. Kim, "Video recognition of simple mastoidectomy using convolutional neural networks: Detection and segmentation of surgical tools and anatomical regions," Comput Methods Programs Biomed, vol. 208, 2021, doi: 10.1016/j.cmpb.2021.106251.

[58] M. Kalia, T. A. Aleef, N. Navab, P. Black, and S. E. Salcudean, "Co-generation and Segmentation for Generalized Surgical Instrument Segmentation on Unlabelled Data," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2021. doi: 10.1007/978-3-030-87202-1_39.

[59] J. Wang, Y. Jin, L. Wang, S. Cai, P. A. Heng, and J. Qin, "Efficient Global-Local Memory for Real-Time Instrument Segmentation of Robotic Surgical Video," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2021. doi: 10.1007/978-3-030-87202-1_33.

[60] K. Zinchenko and K. T. Song, "Autonomous Endoscope Robot Positioning Using Instrument Segmentation with Virtual Reality Visualization," IEEE Access, vol. 9, 2021, doi: 10.1109/ACCESS.2021.3079427.

[61] A. Boonkong, D. Hormdee, S. Sonsilphong, and K. Khampitak, "Surgical Instrument Detection for Laparoscopic Surgery using Deep Learning," in 19th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology, ECTI-CON 2022, 2022. doi: 10.1109/ECTI-CON54298.2022.9795561.

[62] Z. L. Ni et al., "SurgiNet: Pyramid Attention Aggregation and Class-wise Self-Distillation for Surgical Instrument Segmentation," Med Image Anal, vol. 76, 2022, doi: 10.1016/j.media.2021.102310.

[63] X. Sun, Y. Zou, S. Wang, H. Su, and B. Guan, "A parallel network utilizing local features and global representations for segmentation of surgical instruments," Int J Comput Assist Radiol Surg, vol. 17, no. 10, 2022, doi: 10.1007/s11548-022-02687-z.

[64] P. F. Baldi, S. Abdelkarim, J. Liu, J. K. To, M. D. Ibarra, and A. W. Browne, "Vitreoretinal Surgical Instrument Tracking in Three Dimensions Using Deep Learning," Transl Vis Sci Technol, vol. 12, no. 1, 2023, doi: 10.1167/tvst.12.1.20.

[65] J. N. Paranjape, N. G. Nair, S. Sikder, S. S. Vedula, and V. M. Patel, "AdaptiveSAM: Towards Efficient Tuning of SAM for Surgical Scene Segmentation," Aug. 2023, [Online]. Available: http://arxiv.org/abs/2308.03726

Bioinformatics), 2015. doi: 10.1007/978-3-319-24574-4_28.

[47] V. I. Iglovikov and A. A. Shvets, "TernausNet," in Computer-Aided Analysis of Gastrointestinal Videos, 2021. doi: 10.1007/978-3-030-64340-9_15.

[48] A. Chaurasia and E. Culurciello, "LinkNet: Exploiting encoder representations for efficient semantic segmentation," in 2017 IEEE Visual Communications and Image Processing, VCIP 2017, 2018. doi: 10.1109/VCIP.2017.8305148.

[49] D. Pakhomov and N. Navab, "Searching for Efficient Architecture for Instrument Segmentation in Robotic Surgery," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2020. doi: 10.1007/978-3-030-59716-0_62.

[50] S. M. Kamrul Hasan and C. A. Linte, "U-NetPlus: A Modified Encoder-Decoder U-Net Architecture for Semantic and Instance Segmentation of Surgical Instruments from Laparoscopic Images," in Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS, 2019. doi: 10.1109/EMBC.2019.8856791.

[51] Z. L. Ni, G. Bin Bian, Z. G. Hou, X. H. Zhou, X. L. Xie, and Z. Li, "Attention-Guided Lightweight Network for Real-Time Segmentation of Robotic Surgical Instruments," in Proceedings - IEEE International Conference on Robotics and Automation, 2020. doi: 10.1109/ICRA40945.2020.9197425.

[52] M. Islam, V. VS, C. M. Lim, and H. Ren, "ST-MTL: Spatio-Temporal multitask learning model to predict scanpath while tracking instruments in robotic surgery," Med Image Anal, vol. 67, 2021, doi: 10.1016/j.media.2020.101837.

[53] A. Nazir et al., "SPST-CNN: Spatial pyramid based searching and tagging of liver's intraoperative live views via CNN for minimal invasive surgery," J Biomed Inform, vol. 106, 2020, doi: 10.1016/j.jbi.2020.103430.

[54] Y. Fu et al., "More unlabelled data or label more data? a study on semi-supervised laparoscopic image segmentation," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2019. doi: 10.1007/978-3-030-33391-1_20.

[55] S. M. Cho, Y. G. Kim, J. Jeong, I. Kim, H. jin Lee, and N. Kim, "Automatic tip detection of surgical instruments in biportal endoscopic spine surgery," Comput Biol Med, vol. 133, 2021, doi: 10.1016/j.compbiomed.2021.104384.

[56] T. Cheng et al., "Deep learning assisted robotic magnetic anchored and guided endoscope for real-time instrument tracking," IEEE Robot Autom Lett,

[66] A. Kirillov et al., "Segment Anything," Apr. 2023, [Online]. Available: http://arxiv.org/abs/2304.02643

[67] A. Wang, M. Islam, M. Xu, Y. Zhang, and H. Ren, "SAM Meets Robotic Surgery: An Empirical Study on Generalization, Robustness and Adaptation," Aug. 2023, [Online]. Available: http://arxiv.org/abs/2308.07156

[68] S. Petscharnig and K. Schöffmann, "Deep learning for shot classification in gynecologic surgery videos," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2017. doi: 10.1007/978-3-319-51811-4_57.

[69] S. Petscharnig and K. Schöffmann, "Learning laparoscopic video shot classification for gynecological surgery," Multimed Tools Appl, vol. 77, no. 7, 2018, doi: 10.1007/s11042-017-4699-5.

[70]S. Petscharnig, K. Schoffmann, J. Benois-Pineau, S. Chaabouni, and J. Keckstein, "Early and Late Fusion of Temporal Information for Classification of Surgical Actions in Laparoscopic Gynecology," in Proceedings - IEEE Symposium on Computer-Based Medical Systems, 2018. doi: 10.1109/CBMS.2018.00071.

[71] Y. Jin et al., "Multi-task recurrent convolutional network with correlation loss for surgical video analysis," Med Image Anal, vol. 59, 2020, doi: 10.1016/j.media.2019.101572.

[72] Y. Jin et al., "SV-RCNet: Workflow recognition from surgical videos using recurrent convolutional network," IEEE Trans Med Imaging, vol. 37, no. 5, 2018, doi: 10.1109/TMI.2017.2787657.

[73] Y. Liu et al., "LoViT: Long Video Transformer for Surgical Phase Recognition," May 2023, [Online]. Available: http://arxiv.org/abs/2305.08989

[74] I. Funke, A. Jenke, S. T. Mees, J. Weitz, S. Speidel, and S. Bodenstedt, "Temporal coherence-based self-supervised learning for laparoscopic workflow analysis," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2018. doi: 10.1007/978-3-030-01201-4_11.

[75] D. R. Chittajallu et al., "XAI-CBIR: Explainable ai system for content based retrieval of video frames from minimally invasive surgery videos," in Proceedings - International Symposium on Biomedical Imaging, 2019. doi: 10.1109/ISBI.2019.8759428.

[76] B. Namazi, G. Sankaranarayanan, and V. Devarajan, "Attention-based surgical phase boundaries detection in laparoscopic videos," in Proceedings - 6th Annual Conference on Computational Science and Computational Intelligence, CSCI 2019, 2019. doi: 10.1109/CSCI49370.2019.00109.

[77] S. Kannan, G. Yengera, D. Mutter, J. Marescaux, and N. Padoy, "Future-State Predicting LSTM for Early Surgery Type Recognition," IEEE Trans Med Imaging, vol. 39, no. 3, 2020, doi: 10.1109/TMI.2019.2931158.

[78] Y. Chen, Q. L. Sun, and K. Zhong, "Semi-supervised spatio-temporal CNN for recognition of surgical workflow," EURASIP J Image Video Process, vol. 2018, no. 1, 2018, doi: 10.1186/s13640-018-0316-4.

[79] I. Funke, D. Rivoir, and S. Speidel, "Metrics Matter in Surgical Phase Recognition," May 2023, [Online]. Available: http://arxiv.org/abs/2305.13961

[80] I. Funke, S. Bodenstedt, F. Oehme, F. von Bechtolsheim, J. Weitz, and S. Speidel, "Using 3D Convolutional Neural Networks to Learn Spatiotemporal Features for Automatic Surgical Gesture Recognition in Video," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2019. doi: 10.1007/978-3-030-32254-0_52.

[81] T. Khatibi and P. Dezyani, "Proposing novel methods for gynecologic surgical action recognition on laparoscopic videos," Multimed Tools Appl, vol. 79, no. 41–42, 2020, doi: 10.1007/s11042-020-09540-y.

[82] F. Luongo, R. Hakim, J. H. Nguyen, A. Anandkumar, and A. J. Hung, "Deep learning-based computer vision to recognize and classify suturing gestures in robot-assisted surgery," Surgery (United States), vol. 169, no. 5, 2021, doi: 10.1016/j.surg.2020.08.016.

[83] C. I. Nwoye et al., "Recognition of Instrument-Tissue Interactions in Endoscopic Videos via Action Triplets," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2020. doi: 10.1007/978-3-030-59716-0_35.

[84] A. Murali et al., "TSC-DL: Unsupervised trajectory segmentation of multi-modal surgical demonstrations with Deep Learning," in Proceedings - IEEE International Conference on Robotics and Automation, 2016. doi: 10.1109/ICRA.2016.7487607.

[85] J. Xie, H. Zhao, Z. Shao, Z. Shi, and Y. Guan, "A Fast Approach for Multi-Modality Surgical Trajectory Segmentation with Unsupervised Deep Learning," Jiqiren/Robot, vol. 41, no. 3, 2019, doi: 10.13973/j.cnki.robot.180387.

[86] H. Zhao, J. Xie, Z. Shao, Y. Qu, Y. Guan, and J. Tan, "A fast unsupervised approach for multi-modality surgical trajectory segmentation," IEEE Access, vol. 6, 2018, doi: 10.1109/ACCESS.2018.2872635.

[87] R. Tao, X. Zou, and G. Zheng, "LAST: LAtent Space-constrained Transformers for Automatic Surgical Phase Recognition and Tool Presence Detection," IEEE Trans Med Imaging, 2023, doi: 10.1109/TMI.2023.3279838.

[88] D. Kiyasseh et al., "A vision transformer for decoding surgeon activity from surgical videos," Nat Biomed Eng, 2023, doi: 10.1038/s41551-023-01010-8.

[89] B. Zhang et al., "Surgical workflow recognition with temporal convolution and transformer for action segmentation," Int J Comput Assist Radiol Surg, vol. 18, no. 4, 2023, doi: 10.1007/s11548-022-02811-z.

[90] K. Lam et al., "Machine learning for technical skill assessment in surgery: a systematic review," npj Digital Medicine, vol. 5, no. 1. 2022. doi: 10.1038/s41746-022-00566-0.

[91] L. Maier-Hein et al., "Surgical data science for next-generation interventions," Nature Biomedical Engineering, vol. 1, no. 9. 2017. doi: 10.1038/s41551-017-0132-7.

[92] L. Wang et al., "Temporal Segment Networks for Action Recognition in Videos," IEEE Trans Pattern Anal Mach Intell, vol. 41, no. 11, 2019, doi: 10.1109/TPAMI.2018.2868668.

[93] L. Wang et al., "Temporal segment networks: Towards good practices for deep action recognition," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2016. doi: 10.1007/978-3-319-46484-8_2.

[94] I. Funke, S. T. Mees, J. Weitz, and S. Speidel, "Video-based surgical skill assessment using 3D convolutional neural networks," Int J Comput Assist Radiol Surg, vol. 14, no. 7, 2019, doi: 10.1007/s11548-019-01995-1.

[95] Y. Jia et al., "C3D: Generic Features for Video Analysis (Learning Spatiotemporal Features with 3D Convolutional Networks)," 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014.

[96] S. Ji, W. Xu, M. Yang, and K. Yu, "3D Convolutional neural networks for human action recognition," IEEE Trans Pattern Anal Mach Intell, vol. 35, no. 1, 2013, doi: 10.1109/TPAMI.2012.59.

[97] J. Carreira and A. Zisserman, "Quo Vadis, action recognition? A new model and the kinetics dataset," in Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, 2017. doi: 10.1109/CVPR.2017.502.

[98] J. E. Kim, P. Weber, and A. Szabo, "Medical malpractice claims related to cataract surgery complicated by retained lens fragments (an American ophthalmological society thesis)," Trans Am Ophthalmol Soc, vol. 110, 2012.

[99] D. Kitaguchi, N. Takeshita, H. Matsuzaki, T. Igaki, H. Hasegawa, and M. Ito, "Development and Validation of a 3-Dimensional Convolutional Neural Network for Automatic Surgical Skill Assessment Based on Spatiotemporal Video Analysis," JAMA Netw Open, vol. 4, no. 8, 2021, doi: 10.1001/jamanetworkopen.2021.20786.

[100] H. Doughty, D. Damen, and 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, 2018. doi: 10.1109/CVPR.2018.00634.

[101] D. Liu, T. Jiang, Y. Wang, R. Miao, F. Shan, and Z. Li, "Clearness of operating field: a surrogate for surgical skills on in vivo clinical data," Int J Comput Assist Radiol Surg, vol. 15, no. 11, 2020, doi: 10.1007/s11548-020-02267-z.

[102] T. Wang, Y. Wang, and M. Li, "Towards Accurate and Interpretable Surgical Skill Assessment: A Video-Based Method Incorporating Recognized Surgical Gestures and Skill Levels," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2020. doi: 10.1007/978-3-030-59716-0_64.

[103] Z. Li, L. Gu, W. Wang, R. Nakamura, and Y. Sato, "Surgical Skill Assessment via Video Semantic Aggregation," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2022. doi: 10.1007/978-3-031-16449-1_39.

[104] A. Soleymani, A. A. Sadat Asl, M. Yeganejou, S. Dick, M. Tavakoli, and X. Li, "Surgical Skill Evaluation from Robot-Assisted Surgery Recordings," in 2021 International Symposium on Medical Robotics, ISMR 2021, 2021. doi: 10.1109/ISMR48346.2021.9661527.

[105] M. Shafiq and Z. Gu, "Deep Residual Learning for Image Recognition: A Survey," Applied Sciences (Switzerland), vol. 12, no. 18. 2022. doi: 10.3390/app12188972.

[106] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, 2017. doi: 10.1109/CVPR.2017.195.

[107] Y. Gu et al., "Construction of Quantitative Indexes for Cataract Surgery Evaluation Based on Deep Learning," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2020. doi: 10.1007/978-3-030-63419-3_20.

[108] T. S. Kim, M. O'Brien, S. Zafar, G. D. Hager, S. Sikder, and S. S. Vedula, "Objective assessment of intraoperative technical skill in capsulorhexis

using videos of cataract surgery," Int J Comput Assist Radiol Surg, vol. 14, no. 6, 2019, doi: 10.1007/s11548-019-01956-8.

[109] F. Setti et al., "A Multirobots Teleoperated Platform for Artificial Intelligence Training Data Collection in Minimally Invasive Surgery," in 2019 International Symposium on Medical Robotics, ISMR 2019, 2019. doi: 10.1109/ISMR.2019.8710209.

[110] S. S. Vedula, M. Ishii, and G. D. Hager, "Objective Assessment of Surgical Technical Skill and Competency in the Operating Room," Annu Rev Biomed Eng, vol. 19, 2017, doi: 10.1146/annurev-bioeng-071516-044435.