Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Review

Vision-based human action quality assessment: A systematic review

Jiang Liu [a], Huasheng Wang [a,*], Katarzyna Stawarz [a], Shiyin Li [b], Yao Fu [c], Hantao Liu [a]

[a] School of Computer Science and Informatics, Cardiff University, Cardiff, UK
[b] School of Information and Control Engineering, China University of Mining and Technology, Xuzhou, China
[c] School of Engineering, Cardiff University, Cardiff, UK

ARTICLE INFO

Keywords:
Action quality assessment
Computer vision
Human action evaluation
Deep learning

ABSTRACT

Human Action Quality Assessment (AQA), which aims to automatically evaluate the performance of actions executed by humans, is an emerging field of human action analysis. Although many review articles have been conducted for human action analysis fields such as action recognition and action prediction, there is a lack of up-to-date and systematic reviews related to AQA. This paper aims to provide a systematic literature review of existing papers on vision-based human AQA. This systematic review was rigorously conducted following the PRISMA guideline through the databases of Scopus, IEEE Xplore, and Web of Science in July 2024. Ninety-six research articles were selected for the final analysis after applying inclusion and exclusion criteria. This review presents an overview of various aspects of AQA, including existing applications, data acquisition methods, public datasets, state-of-the-art methods and evaluation metrics. We observe an increase in the number of studies in AQA since 2019, primarily due to the advent of deep learning methods and motion capture devices. We categorize these AQA methods into skeleton-based and video-based methods based on the data modality used. There are different evaluation metrics for various AQA tasks. SRC is the most commonly used evaluation metric, with fifty-six out of ninety-six selected papers using it to evaluate their models. Sports event scoring, surgical skill evaluation and rehabilitation assessment are the most popular three scenarios in this direction based on existing papers and there are more new scenarios being explored such as piano skill assessment. Furthermore, the existing challenges and future research directions are provided, which can be a helpful guide for researchers to explore AQA.

Contents

* Corresponding author.
E-mail addresses: liuj137@cardiff.ac.uk (J. Liu), wanghs@cardiff.ac.uk (H. Wang), stawarzk@cardiff.ac.uk (K. Stawarz), lishiyin@cumt.edu.cn (S. Li), fuy49@cardiff.ac.uk (Y. Fu), liuh35@cardiff.ac.uk (H. Liu).

https://doi.org/10.1016/j.eswa.2024.125642
Received 2 November 2023; Received in revised form 17 October 2024; Accepted 26 October 2024
Available online 14 November 2024

1. Introduction

Action Quality Assessment (AQA), which aims to use computer to automatically evaluate the quality of sequences of movements executed by individuals, has attracted growing attention in the computer vision community (Tang et al., 2020). Conventionally, the task of assessing human action quality relies on the professionals observing the complete motion and evaluating it based on established standards. However, manual AQA faces some challenges in real life (Gharasuie, Jennings, & Jain, 2021; Machlin, Chevan, Yu, & Zodet, 2011; MacMahon et al., 2014; Sardari et al., 2023, 2023; Zhou, Feng, Chen, Ban & Pan, 2023). For instance, in a regular rehabilitation training process, physical therapists need to spend a considerable amount of time monitoring a patient's movements during exercise and providing corrective feedback (Sardari et al., 2023). However, even in some developed countries such as the United States and Australia, there are only about 300 trained therapists in a million individuals, causing a shortage of available personnel for providing manual guidance in rehabilitation services (Bettger et al., 2019). In addition, it is expensive to recruit professionals for long-term fitness or rehabilitation training, so not everyone can afford these costs (Gharasuie et al., 2021; Machlin et al., 2011). Furthermore, in some competitive sport games, referees are required to make decisions about the performance of athletes in a limited time under a dynamic environment (MacMahon et al., 2014). It is challenging to manually provide objective and fair scores during a whole game without any mistakes to avoid judging scandals (Parmar & Morris, 2019a; Parmar & Tran Morris, 2017). Thus, there is an urgent need to develop technological systems to help humans evaluate the action quality. In recent years, a lot of attempts have been made to explore using Artificial Intelligence (AI) to assess human actions. For example, the International Football Association Board started to introduce Video Assistant Referees (VAR) to assist make decisions in the 2018 Fédération Internationale de Football Association (FIFA) World Cup (Brunnström et al., 2023). In addition, an AI-based scoring system was used in 2019 World Artistic Gymnastic Championships for pommel horse, still rings and vault events (Jakab, Davis, & Whyte, 2023). As the key technology of home-based rehabilitation (Baca, Dabnichki, Hu, Kornfeind, & Exel, 2022; Li, Hu et al., 2022; Liao, Vakanski, Xian, Paul, & Baker, 2020), AI coach (Liao, Hwang, Wu, & Koike, 2023; Toshniwal, Patil, & Vachhani, 2022) and VAR, the development of AQA system is urgently required. It can play a critical role in providing individuals with feedback on their executed exercises, enabling them to adjust their movement in time to maximize the benefits of training and reduce the risk of injuries in the absence of professionals. Also, this could aid professionals such as therapists and referees to make their decision and reduce the workload burden.

Due to the diversity and complexity of human actions, the task of AQA is still challenging and needs to be paid more effort. In contrast to another popular field of human action analysis, Human Action Recognition (HAR), which focuses on classifying different coarse-granularity features among various actions, AQA needs to discriminate fine-granularity internal differences within a specific action, which requires better perception ability (Wang, Yang, Zhai, Yu et al., 2021). Additionally, while recognizing an action may only require analysing a portion of the action (Karpathy et al., 2014), assessing the quality of an action need to analyse the whole action sequence (Parmar & Tran Morris, 2017), which contains richer and more complex information.

In recent years, Scholars have used various modalities for human activity analysis (Gao, Cui et al., 2023; Setiawan, Yahya, Chun, & Lee, 2022; Su et al., 2022; Sun et al., 2022; Yin, He, Soomro, & Yuan, 2023; Zhu, Lu, Gan, & Hou, 2021). It can be roughly divided into two categories: visual modalities and non-visual modalities. Generally, Red, Green, Blue (RGB) videos/images, depth videos/images, skeleton data sequences and infrared sequences are considered visual modalities (Karayaneva, Sharifzadeh, Jing, Chetty, & Tan, 2019; Kong & Fu, 2022), as they can obtain rich appearance or posture information for visually presenting human action. In contrast, wearable sensors (Kim, Lee, & Hong, 2023; Sigcha et al., 2023), audio (Do, Welch, & Sheng, 2021), radar (Li, He & Jing, 2019) and Wifi (Hao, Shi, & Liu, 2022) can collect non-visual information to present human behaviours, which are also used for HAR in situations where privacy protection is required. However, wearable-based methods can be invasive and cause discomfort during daily activities. Audio- and WiFi-based methods may lack robustness in complex environments and be difficult to capture subtle internal action differences information to fulfil the requirements of AQA. Radar-based methods can be costly. In contrast, Cameras are ubiquitous tools of low cost in our life and can directly obtain subtle appearance (Chen et al., 2020). Therefore, we only focus on vision-based methods for human AQA in this paper.

1.1. Literature survey

While quite a few reviews have been published to summarize the advancement of HAR (Al-Faris, Chiverton, Ndzi, & Ahmed, 2020; Islam, Nooruddin, Karray, & Muhammad, 2022; Majumder & Kehtarnavaz, 2021; Muhamada & Mohammed, 2021; Sun, Ke et al., 2023), very few reviews have been published to analyse the AQA. For example, Ahad, Antar, and Shahid (2019) provided a short review of the computer vision-based action understanding in assistive healthcare. They reviewed available sensing devices and benchmark datasets, as well as summarized the challenges and difficulties in the automation of assistive healthcare system. However, the articles relevant to AQA in this review were limited in number and scope, as they primarily concentrated on HAR in healthcare and rehabilitation applications.

In contrast, (Lei, Du, Zhang, Ye, & Chen, 2019) provided a clear definition of human action evaluation and clarified the differences from action recognition and action prediction. Additionally, they conducted a comprehensive survey of human AQA methods, benchmark datasets and evaluation criteria in various fields, including healthcare, physical rehabilitation, skill training, and sports activity scoring. However, since the publication of this article in 2019, significant progress has been made in the field of AQA over the past five years, which was not included in the article.

Finally, Wang, Yang, Zhai, Yu et al. (2021) summarized the existing video-based AQA datasets and models in sports and medical care and discussed the challenges and future development direction. However, there was limited attention given to models based on skeleton data, which has been conducted in many studies in this field of research, e.g. Huang, Yang, Luo, and Zhang (2023) and Liang, Luo, Gao, and Lu (2021).

1.2. Motivation

AQA has immense potential for application in various real-world scenarios requiring movement quality assessment. For example, in healthcare, AQA forms the foundation for remote rehabilitation solutions. It assists rehabilitation therapists in evaluating the movement quality of patients, thereby reducing their workload. Additionally, in medicine AQA can evaluate surgical skills, thus lowering the cost of training specialized surgeons. In the fitness domain, AQA can help trainees improve their exercise quality without the supervision of a personal trainer, reducing the risk of injury. In sports, AQA can support referees in scoring, alleviating their pressure and maintaining fairness and objectivity. Therefore, AQA has vast prospects and is worth further research.

In recent years, although many methods and a wide range of applications have been proposed for AQA, there is a lack of up-to-date and systematic reviews related to AQA. This paper aims to investigate the current state of AQA, including its applications, data acquisition techniques, available datasets, methods, evaluation metrics, as well as the challenges and future directions in the field. We aim to offer valuable insights and a foundational reference for researchers and practitioners, driving further innovation and development in AQA.

1.3. Contribution

The contributions of this paper are as follows:

- To the best of our knowledge, this study represents the first systematic literature review to investigate up-to-date research in vision-based AQA. It provides a comprehensive synthesis of current advancements and emerging trends in the field.
- This study offers a detailed examination of 96 papers, including their applications, datasets, data modalities, methods, and evaluation metrics. We conducted a rigorous statistical analysis and categorization based on these aspects. This review can serve as a guideline for new researchers entering the field.
- This review identifies current challenges in existing research, providing valuable insights and recommendations for future studies. These suggestions are intended to inspire the development of new methods and applications within the AQA field.

Table 1
Databases and search terms employed in this article.

Sources	Keywords
Scopus	action quality assessment
IEEE Xplore	action assessment
Web of Science	action quality evaluation
	human action evaluation
	movement quality assessment
Search String: TITLE-ABS-KEY (“action quality assessment” OR “action assessment” OR “action quality evaluation” OR “human action evaluation” OR “movement quality assessment”) AND (LIMIT-TO (DOCTYPE , “ar”) OR LIMIT-TO (DOCTYPE , “cp”)) AND (LIMIT-TO (SUBJAREA , “COMP”)) AND (LIMIT-TO (LANGUAGE , “English”))	

The remainder of our paper is structured as follows: Section 2 shows the research methodology. The research question, search strategy and selection process are defined in this section. In Section 3, application, data acquisition, datasets, methods, and metrics are reviewed based on selected literature. The analysis and findings based on our research results are presented in Section 4. Section 5 shows current challenges and future research in AQA field. The conclusion is provided in Section 6.

2. Research methodology

This review was performed according to the systematic literature review process proposed by Kitchenham, Charters, et al. (2007) and the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) updated guideline for reporting systematic reviews (Page et al., 2021).

2.1. Research question

This article aims to review current literature relevant to vision-based human AQA and provide a comprehensive understanding of application, equipment, datasets, models and evaluation metrics. In particular, the following research questions were defined:

1. What are the existing applications of AQA?
2. What are the data acquisition methods for AQA?
3. What existing datasets have been used for AQA?
4. What are the methods used for AQA?
5. What are the most common metrics used to evaluate AQA model?

2.2. Search strategy

To conduct the systematic review, keywords and boolean operators were utilized to create the search string. “Action quality assessment” and some synonyms were defined as key items to search relevant articles (see Table 1). Please note that we did not use any keywords associated with vision since a variety of words can represent vision, such as RGB, video and some names of optical equipment. Scopus, IEEE Explore and Web of Science were selected as the databases. In addition, we did not set time constraints in order to produce a comprehensive review including earlier research. The search was conducted on 16 July 2024 and the search strategy was utilized across the title, abstract and keywords of articles within the databases to find articles related to the topic. Table 1 shows the databases, the keywords and an example of the search string.

Studies were included if they satisfied the following inclusion criteria: (1) published in a journal or conference proceedings, (2) related to computer science, (3) published in English. Additionally, the following exclusion criteria were applied: (1) duplicated articles, (2) articles that presented results of surveys or reviews, (3) articles not related to using vision-based methods, and (4) articles not related to human AQA.

Table 2
Overview of existing applications in AQA.

Application	Numbers	Related paper
Sports event scoring	53	Bai et al. (2022), Dadashzadeh, Duan, Whone, and Mirmehdi (2024), Dong et al. (2021), Du, He, Wang, and Wang (2023), Fang, Zhou and Li (2023), Farabi et al. (2022), Gan et al. (2024), Gao, Pan, Zhang and Zheng (2023), Gedamu, Ji, Yang, Shao, and Shen (2023), He et al. (2024), Hirosawa, Kato, Yamashita, and Aoki (2023), Huang and Li (2024), Ji et al. (2023), Ke, Xu, Lin, and Guo (2024), Lei, Li, Zhang, Du, and Gao (2023), Lei, Zhang, and Du (2021), Lei, Zhang, Du, Hsiao, and Chen (2020), Li, Chai, and Chen (2018), Li, Chai and Chen (2019), Li, Cui, Kitahara and Sagawa (2022), Li, Lei, Zhang and Du (2021), Li, Lei, Zhang, Du and Gao (2022), Li, Zhang, Dong, Lei and Du (2022), Li, Zhang, Dong, Lei and Du (2023), Li, Zhang, Lei et al. (2022), Lian and Shao (2023), Liu, Cheng and Ikenaga (2023), Liu, Zhai, Zheng and Fang (2023), Matsuyama, Kawaguchi, and Lim (2023), Nagai, Takeda, Matsumura, Shimizu, and Yamamoto (2021), Nekoui, Cruz, and Cheng (2021), Pan, Gao, and Zheng (2019, 2022), Parmar and Morris (2019a, 2019b), Parmar and Tran Morris (2017), Pirsiavash, Vondrick, and Torralba (2014), Roditakis, Makris, and Argyros (2021), Sun, Hu et al. (2023), Tang et al. (2020), Wang, Du, Li and Wang (2020), Wang, Yang, Zhai, Chen and Zhang (2021), Xiang, Tian, Reiter, Hager, and Tran (2018), Xu, Rao et al. (2022), Xu, Zeng and Zheng (2022), Yu, Rao, Zhao, Lu and Zhou (2021), Zeng et al. (2020), Zeng and Zheng (2024), Zhang, Chen et al. (2024), Zhang, Dai et al. (2023), Zhang, Dong, Lei, Yang and Du (2022, 2023), Zhang, Pan, Gao and Zheng (2022, 2024), Zhang, Wang, Zhuang and Wang (2023), Zhang, Xiong and Mi (2022) and Zhou, Ma, Shum and Liang (2023)
Surgical skill evaluation	13	Baby et al. (2022), Dadashzadeh et al. (2024), Gao, Pan et al. (2023), Gao et al. (2020), Ke et al. (2024), Pan et al. (2019, 2022), Sun, Hu et al. (2023), Tang et al. (2020), Yu, Rao et al. (2021), Zhang, Chen et al. (2024), Zhang, Pan et al. (2024) and Zhou, Ma et al. (2023)
Rehabilitation assessment	13	Dadashzadeh et al. (2024), Fang, Luo et al. (2023), Kanade, Sharma, and Muniyandi (2023a, 2023b), Li, Ling and Xia (2023), Mourchid and Slama (2023), Sardari et al. (2024), Venkataraman and Turaga (2016), Venkataraman et al. (2016, 2013, 2014), Yu, Liu, and Chan (2020) and Yu, Liu, Chan, and Chen (2024)
Physical exercise assessment	3	Çeliktutan, Akgül, Wolf, and Sankur (2013), Chariar, Rao, Irani, Suresh, and Asha (2023) and Dajime, Smith, and Zhang (2020)
Fitness action assessment	4	Jin et al. (2016), Joung, Byun, and Baek (2023), Li, Hu, Guo, Wang and Shen (2021) and Wang et al. (2023)
Martial arts assessment	4	Li, Hu et al. (2022), Li, Tian and Li (2023), Wang, Li, and Hu (2022) and Yuan (2024)
Behaviour therapy	5	Li, Bhat and Barmaki (2021), Li, Chheang et al. (2023), Yu, Liu, Chan, Yang and Wang (2021), Zhang, Zhou and Liu (2023) and Zhou, Cai et al. (2023)
Piano skill assessment	1	Parmar, Reddy, and Morris (2021)
Hand skill evaluation	2	Wang, Jin, Wang, Wang and Li (2020) and Zhang, Pan et al. (2024)
Golf skill assessment	1	Ingwersen et al. (2023)
Pull-ups test	1	Liu, Wang et al. (2023)
Running performance analysis	2	Freire-Obregón, Lorenzo-Navarro, and Castrillón-Santana (2022) and Freire-Obregon, Lorenzo-Navarro, Santana, Hernandez-Sosa, and Castrillon-Santana (2023)
Daily action assessment	1	Gao, Pan et al. (2023)
Windsurfing assessment	1	Nagai, Takeda, Suzuki, and Seshimo (2024)
Dance assessment	1	Hipiny, Ujir, Alias, Shanat, and Ishak (2023)

2.3. Selection process

Firstly, the primary studies were identified according to the search strategy mentioned before. Next, articles were screened by their title, abstract and keywords and irrelevant articles were removed. In addition, the full text of the articles was analysed to check if it meets the eligibility. Finally, the references of the selected articles were reviewed in order to identify additional relevant studies.

In total, 230 articles were retrieved from the databases by applying the search strategy. After removing duplicates and irrelevant articles by applying the criteria, 94 research articles remained. Next, the references of the selected papers were reviewed to identify additional relevant studies, resulting in a final list of 96 papers for the systematic literature review. Fig. 1 illustrates the entire procedure of identification, screening, eligibility assessment and inclusion for the final list of papers. Fig. 2 shows the yearly distribution of included 96 articles published in the field of AQA from 2013 to 2024. It is worth noting that due to this systematic review being conducted in July 2024, the number of articles published in 2024 is limited (only 13). We can observe that there has been a significant increase in the number of publications since 2019 with the growing attention in AQA.

3. Results

A total of 96 peer-reviewed research papers on AQA were studied. This section presents a systematic view of the application, data acquisition, datasets, methods and evaluation metrics mentioned in the selected papers, organized according to the research questions defined in Section 2. A summary of the 96 papers is presented in Table 13 in Appendix.

3.1. Application

With the prosperity and development of motion capture devices and deep learning methods, scholars have tried to apply AQA to various real-world scenarios, including sports event scoring, surgical skill evaluation, rehabilitation assessment, physical exercise assessment, fitness action assessment, Tai Chi Quan gesture assessment, social behaviour analysis, piano skill evaluation, hand skill evaluation, pull-up test, running performance analysis and daily action assessment. In Table 2, we summarize the existing applications of AQA and the number of identified papers. It should be noted that one AQA method may be applied to multiple fields in a study. For example, Yu, Rao et al. (2021) proposed a contrastive regression framework and applied it to sports event scoring and surgical skill evaluation. We observed the

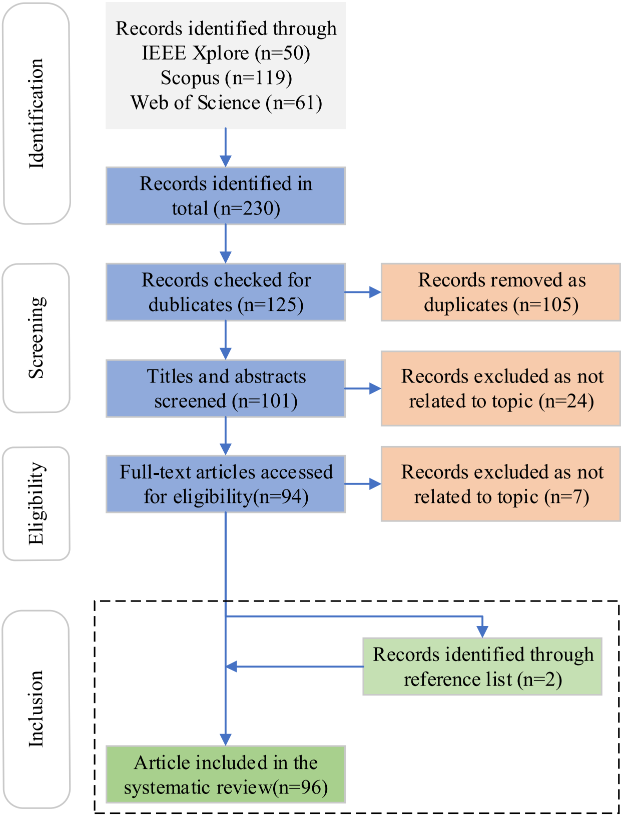
Identification
Screening
Eligibility
Inclusion
Records identified through
IEEE Xplore (n=50)
Scopus (n=119)
Web of Science (n=61)
Records identified in total (n=230)
Records checked for dublicates (n=125)
Records removed as duplicates (n=105)
Titles and abstracts screened (n=101)
Records excluded as not related to topic (n=24)
Full-text articles accessed for eligibility(n=94)
Records excluded as not related to topic (n=7)
Records identified through reference list (n=2)
Article included in the systematic review(n=96)

Fig. 1. Systematic literature review process.

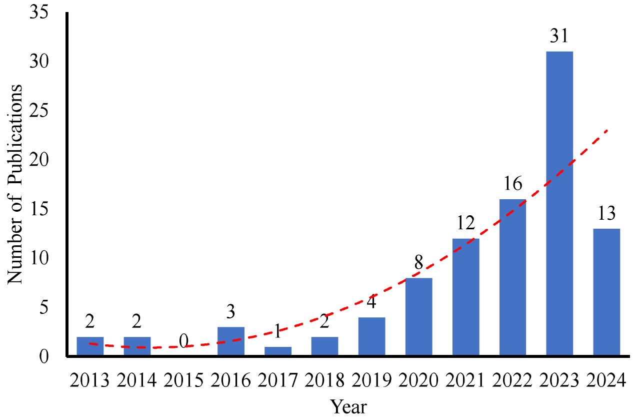

Fig. 2. The number of publications in AQA by year of publication. The red dashed line shows the fit of the number of publications from 2013 to 2024.

number of papers related to sports event scoring far exceeds those in other domains. This is because there are some publicly available datasets, which can promote the development. The top 3 common existing applications in AQA are sports event scoring (53 identified papers), surgical skill evaluation (13) and rehabilitation assessment (13). Furthermore, as shown in Table 2, AQA has been applied to some new fields, such as piano skill assessment.

3.2. Data acquisition

As mentioned before, there are several vision-based data modalities, such as RGB, depth, skeleton data and infrared sequences. However, among the papers we have identified, RGB video information and skeleton information are the most commonly used data to evaluate action quality, as they can provide rich interpretable features. In addition, some other information can be used in conjunction with them such as depth and audio data. It is worth noting that we did not find any research in selected papers that utilized infrared sequences for AQA. Therefore, we grouped data acquisition into RGB video data and skeleton data.

3.2.1. RGB video data

In recent years, with the development of deep learning, many studies try to use convolutional neural networks (CNN) to extract motion features directly from RGB videos. These RGB videos are captured by various cameras, such as professional sports cameras, web cameras and smartphone cameras. Due to the diverse range of available RGB cameras and the limited mention of specific models in the literature, we did not provide a detailed list of all the models.

3.2.2. Skeleton data

There are three main methods to acquire skeleton data: optical motion capture systems, depth cameras and pose estimation algorithms. In Table 3, we summarize skeleton data acquisition in AQA based on the included 96 papers. We identified 6 related papers using optical motion capture systems, 14 using depth cameras and 18 using pose estimation algorithms methods. We described each skeleton data acquisition method below.

Optical motion capture systems. In the earlier stage of vision-based AQA, optical motion capture systems such as Vicon, Qualisys, and Vinci Surgical System were used to obtain the skeleton data. These systems are regarded as the golden standard in optical tracking sensors because of their excellent performance in tracking joints' positions in both static and dynamic scenarios. However, the poor flexibility (multiple cameras must be installed before capturing), high-cost and poor comfort (participants need to wear markers during the capturing process) limit them to a conditioned laboratory with environment settings.

Commercial depth cameras. With the advent of some low-cost commercial depth cameras, more devices are available for AQA research. Microsoft released Kinect v1, Kinect v2 and Kinect Azure in 2012, 2014 and 2019 respectively. Kinect v1 is embedded with a vision RGB camera, a Structured light depth camera, multiple microphones, and a motorized tilt. It can capture resolution of 640 × 480 RGB and 320 × 240 depth images at a frame rate of 30 fps as well as provide 15 or 20 joints' 3D coordinates by Microsoft Software Development Kit (SDK). Kinect v2 adopts a time-of-flight depth sensor instead of SL sensor to improve depth perception ability based on Kinect v1. It can capture a resolution of 1920 × 1080 RGB and 512 × 424 depth images at a frame rate of 30 fps. Azure Kinect, the latest Kinect, is much smaller and lighter than the previous two generations. It can capture a resolution of 3840 × 2160 RGB and a maximum of 1024 × 1024 depth images at a frame rate of 30 fps. In addition, Asus Xtion Pro Live is embedded with a RGB camera, a structured light depth camera 2 microphones, which allows developers to record audio and track 15 body joints in a whole body movement. It can capture a resolution of 1280 × 1024 RGB images at a frame rate of 30 fps and a resolution of 640 × 480 depth images at a frame rate of 30 fps. Table 4 shows the main features of these mentioned commercial depth cameras.

Pose estimation algorithms. The advent of human pose estimation algorithms also enables us to obtain skeleton information more easily, which can detect the keypoints of the human body in RGB videos and provide 3D real-time coordinates. Generally speaking, there are two types of human pose estimation algorithms: top-down methods and bottom-up methods (Lan, Wu, Hu, & Hao, 2023). For top-down methods, each person is first detected and assigned to a separate bounding box. Then, the keypoints are estimated in each bounding

Table 3
Overview of skeleton data acquisition.

Data acquisition	Numbers	Related paper
Pptical motion capture system	6	Venkataraman and Turaga (2016), Venkataraman et al. (2016, 2013, 2014), Yu et al. (2020) and Yu, Liu et al. (2021)
Depth cameras	14	Chariar et al. (2023), Dajime et al. (2020), Fang, Luo et al. (2023), Jin et al. (2016), Kanade et al. (2023a, 2023b), Li, Hu et al. (2022), Li, Ling et al. (2023), Mourchid and Slama (2023), Sardari et al. (2024), Wang et al. (2022), Yu et al. (2020, 2024) and Yu, Liu et al. (2021)
Pose estimation algorithms	18	Gao, Pan et al. (2023), Hipiny et al. (2023), Hirosawa et al. (2023), Joung et al. (2023), Lei et al. (2023, 2020), Li, Chheang et al. (2023), Li, Hu et al. (2021), Li, Lei et al. (2021, 2022), Li, Tian et al. (2023), Liu, Wang et al. (2023), Nekoui et al. (2021), Pirsiavash et al. (2014), Wang, Jin et al. (2020), Wang et al. (2023), Zhang, Wang et al. (2023) and Zhang, Zhou et al. (2023)

Table 4
Overview of RGB-D cameras.

Depth sensors	Year	RGB camera	Depth camera	Measuring range	Field of view	Skeleton joints
Kinect v1	2012	1280 × 960 px at 12 Hz 640 × 480 px at 30 Hz	Structured light 320 × 240 px at 30 Hz	0.8–4 m	H:57°, V:43°	20
Kinect v2	2014	1920 × 1080 px at 30 Hz	Time-of-flight 512 × 424 px at 30 Hz	0.5–4.5 m	H:70°, V:60°	25
Azure Kinect	2019	3840 × 2160 px at 30 Hz	Time-of-flight 640 × 576 px at 30 Hz 512 × 512 px at 30 Hz 1024 × 1024 px at 15 Hz	0.5–3.86 m; 0.5–5.46 m; 0.25–2.88 m; 0.25–2.21 m. (varies in different modes)	H:75°, V:65° H:120°, V:120°	32
Asus Xtion Pro Live	2012	1280 × 1024 px at 30 Hz	Structured light 640 × 480 px at 30 Hz 320 × 240 px at 60 Hz	0.8 m–3.5 m	H:58°, V:45°	15

Table 5
Overview of pose estimation algorithms.

Algorithms	Maximum keypoints	Method	Can detect multi-person
OpenPose (Cao et al., 2017)	135	Bottom-up	Yes
PoseNet (Papandreou et al., 2017)	17	Top-down	Yes
MediaPipe (Bazarevsky et al., 2020)	33	Top-down	No
MoveNet (Jo & Kim, 2022)	17	Bottom-up	Yes

box. For bottom-up methods, all the keypoints of human body are first detected and then grouped into different persons. OpenPose (Cao, Simon, Wei, & Sheikh, 2017), the first open-source 2D multi-person pose estimation algorithm, was released in 2017. It can detect body, hand, foot, and facial keypoints on a single image with a total of 135 keypoints. MoveNet (Jo & Kim, 2022), which was released in 2021, is a pose detection model that detects 17 keypoints of a single person in real-time. MediaPipe (Bazarevsky et al., 2020) was released in 2020 and it can detect a maximum of 33 keypoints (3D landmarks) from an RGB input. PoseNet (Papandreou et al., 2017) was released in 2017 and it can detect 17 keypoints in a single person. Table 5 shows the main features of these pose estimation algorithms.

3.3. Dataset

We reviewed various public vision-based AQA datasets. The majority of them are focused on sports event scoring in Olympic events such as diving, skating and gymnastic vault. Because ground-truth scores can be obtained in these sports through the detailed scoring criteria. In addition, some datasets are related to exercise or rehabilitation movement assessment, including Tai Chi Chuan assessment, gait quality assessment and rehabilitation movement assessment, while others are related to skill rating such as piano skill rating, and surgical skill rating. We can find some researchers try to build AQA dataset in various fields, such as piano skill assessment. The datasets are listed in Table 6 according to published year and described below.

MIT Olympics dataset (Pirsiavash et al., 2014): The MIT Olympics dataset consists of diving and figure skating videos of Olympics and worldwide championships. For diving, a total of 159 videos with slow motion are collected, each is almost 150 frames and captured at a frame rate of 60 fps. In addition, ground truth scores from judges and feedback proposals of actions from professional coaches are provided. For figure skating, 150 videos and judge's score are collected, each is about 4200 frames and captured at a frame rate of 24 fps.

JIGSAW (Gao et al., 2014): The JIGSAW dataset is JHU-ISI Gesture and Skill Assessment Working Set, collected from 8 doctors with different levels of skill for three surgical tasks: suturing, needle-passing and knot-tying. All doctors repeated each task 5 times and the kinematic and video data was captured by the Vinci Surgical System. As a result, 36 samples of knot-tying, 28 samples of needle-passing and 39 samples of suturing were recorded and ground-truth scores from domain experts were provided.

UNLV-vault, skating, diving (Parmar & Tran Morris, 2017): The UNLV-Dive dataset consists of 370 samples, which is extended from MIT-Dive by including more rounds. The UNLV-Vault dataset contains 176 samples with an average length of almost 75 frames. UNLV-Skating includes 171 videos with an average length of 4500 frames.

UMONS-TAICHI (Tits et al., 2018): the UMONS-TAICHI dataset contains 2200 samples of 13 classes of Tai Chi Quan martial art gestures performed by 12 participants with various skill levels. It was collected by 3D motion capture system Qualisys at 179 Hz and Microsoft Kinect V2 at 30 Hz simultaneously. In addition, the levels of participants are evaluated by domain experts on a scale of 0 to 10.

Walking gait dataset (Nguyen et al., 2018): The Walking gait dataset contains 81 samples performed by 9 participants with 9 different levels of walking gates, which is appropriate for gait quality assessment. To create the asymmetry walking gates, each participant was required to walk without a pad and walk with 8 types of different pads under the foot. each sample contains 1200 consecutive frames captured by Kinect V2. point cloud, skeleton, and frontal silhouette data are provided in the dataset.

UI-PRMD dataset (Vakanski et al., 2018): The UI-PRMD dataset contains 10 different movements which are commonly used for patients in physical rehabilitation programs. 10 healthy participants were asked to perform each movement 10 repetitions in both correct and incorrect manners. The angle and position of joints were collected by Kinect v2 and Vicon optical tracking system simultaneously.

Table 6
Overview of publicly available datasets.

Dataset	Scene	Action categories	Samples	Data modality	Website
MIT Olypics (Pirsiavash et al., 2014)	Sports	2	309	video data	kihttps: //redirect.cs.umbc.edu/hpirsiav/quality.html
JIGSAWS (Gao et al., 2014)	Skill	3	103	kinematic data video data	https://cirl.lcsr.jhu.edu/research/hmm/datasets/jigsaws_release/
UNLV-vault (Parmar & Tran Morris, 2017)	Sports	1	370	video data	http://rtis.oit.unlv.edu/datasets.html
UNLV-skating (Parmar & Tran Morris, 2017)	Sports	1	176	video data	http://rtis.oit.unlv.edu/datasets.html
UNLV-diving (Parmar & Tran Morris, 2017)	Sports	1	171	video data	http://rtis.oit.unlv.edu/datasets.html
UMONS-TAICHI (Tits, Laraba, Caulier, Tilmanne, & Dutoit, 2018)	Exercise	13	2200	kinematic data	https://github.com/numediart/UMONS-TAICHI
Walking gait (Nguyen, Huynh, & Meunier, 2018)	Rehabilitation	9	81	kinematic data	http://www-labs.iro.umontreal.ca/labimage/GaitDataset/
UI-PRMD (Vakanski, Jun, Paul, & Baker, 2018)	Rehabilitation	10	100	kinematic data	https://webpages.uidaho.edu/ui-prmd/
AQA-7 (Parmar & Morris, 2019a)	Sports	7	1189	video data	http://rtis.oit.unlv.edu/datasets.html
MTL-AQA (Parmar & Morris, 2019b)	Sports	16	1412	video data	https://github.com/ParitoshParmar/MTL-AQA
FIS-V (Xu et al., 2019)	Sports	1	500	video data	https://github.com/loadder/MS_LSTM
Rhythmic Gymnastics (Zeng et al., 2020)	Sports	4	1000	video data	https://github.com/qinghuannn/ACTION-NET
TASD-2 (Gao et al., 2020)	Sports	2	606	video data	https://www.isee-ai.cn/~gaojibin/ProjectAIM.html
PISA (Parmar et al., 2021)	Skill	1	992	video data	https://github.com/ParitoshParmar/Piano-Skills-Assessment
FineDiving (Xu, Rao et al., 2022)	Sports	1	3000	video data	https://github.com/xujinglin/FineDiving
NETS (Baby et al., 2022)	Skill	1	100	video data	
PaSk (Gao, Pan et al., 2023)	Sports	1	1018	video data	Empty
FineFS (Ji et al., 2023)	Sports	1	1167	kinematic data video data	https://github.com/yanliji/FineFS-dataset
MMASD (Li, Chheang et al., 2023)	Behaviour	11	1315	2D & 3D kinematic data optical flow data	https://github.com/Li-Jicheng
LOGO (Zhang, Dai et al., 2023)	Sports	12	200	video	https://github.com/shiyi-zh0408/LOGO

AQA-7 (Parmar & Morris, 2019a): The AQA-7 dataset contains 1189 samples from seven Olympics events, including singles diving-10 m platform, gymnastic vault, big air skiing, big air snowboarding, synchronous diving-3 m springboard, synchronous diving-10 m platform, and trampoline. Although score rules and ranges are different among these sports, the ground-truth score is provided in the dataset.

MTL-AQA (Parmar & Morris, 2019b): the MTL-AQA dataset, the largest diving dataset to date and the first multitask AQA dataset, was released in 2019. It contains 1412 diving samples collected from 16 different events, including male and female athletes, single or synchronized diving, 3 m springboard and 10 m platform. Furthermore, the AQA score (including final score, difficulty degree and execution score) from judges, diving classes, and text commentary from television analysts are provided to ensure this dataset can be used for multitask learning.

FIS-V (Xu et al., 2019): As a long-term AQA video dataset, the FIS-V contains 500 high-quality figure skating videos from international competitions. Each video is about 4300 frames, captured at a frame rate of 25 fps. In addition, Total Element Score (TES) and Total Program Component Score (PCS) which are given by nine judges to evaluate the performance of the skater at each stage over the whole competition are provided in the dataset.

Rhythmic Gymnastics Dataset (Zeng et al., 2020): As a long-term AQA video dataset, the Rhythmic Gymnastics contains 1000 samples of 4 types of gymnastic routines from the International Rhythmic Gymnastics Competition, including ball, clubs, hoop and ribbon. Each video is about 95 s, captured at a frame rate of 25 fps. Moreover, three types of scores from professional referees are provided: a difficulty score, an execution score and a total score.

TASD-2 (Gao et al., 2020): The TASD-2 dataset contains 606 samples of synchronous diving-3 m springboard, and synchronous diving-10 m platform captured in front view, which can be used for interactive action assessment. Each video was clipped to 102 frames with a format of 320 × 240. In addition, the difficulty score, execution score, synchronization score and final score from professional referees were annotated in the dataset.

PISA (Parmar et al., 2021): The PISA dataset, collected from piano-playing videos on YouTube, is the first piano skill assessment dataset. It contains 992 unique samples and each sample contains 160 frames. Player skill level, song difficulty level, name of the song, and a bounding box around the pianist's hands are provided in the dataset.

FineDiving (Xu, Rao et al., 2022): The Finediving dataset, which is the first fine-grained sports video in AQA, was released in 2022. It contains 3000 samples collected from different diving events such as the Olympics, World Championships, World Cup and European Aquatics Championships on YouTube. Action scores, action types, sub-action types, coarse- and fine-grained temporal boundaries are provided in the dataset.

NETS (Baby et al., 2022): The NETS dataset contains 100 short videos captured from a box trainer, which is used for imparting neuro-endoscope skills. 6 neuro-endoscope experts and 6 trainees were asked to transfer 6 rings in a pre-defined manner using biopsy forceps and an endoscope and videos were captured by a camera at a frame rate of 25 fps. In addition, the skills were evaluated by an expert neurosurgeon on a scale of 1 to 10.

PaSk (Gao, Pan et al., 2023): The PaSk dataset contains 1018 videos collected from pair figure skating events. The length of each video varies between 100 to 1000 frames, depending on the duration of the performance. Furthermore, the dataset includes the final score of the entire performance and sub-scores of every independent action given by referees.

FineFS (Ji et al., 2023): The FineFS dataset contains 1167 skating videos, including 729 short program videos and 438 free skating videos. The length of each video is around 160 s and 240 s for short program and free skating respectively. The frame rate is 25 fps. RGB and Skeleton data are both provided in this dataset. In addition to TES and PCS, more detailed annotations are provided including base value, grade of execution, skating skills, transition, performance, choreography, interpretation and scores from multiple judges.

MMASD (Li, Chheang et al., 2023): The MMASD dataset includes 1315 videos segmented from play therapy intervention recordings of 32 children with ASD. Four types of data are provided in this dataset including optical flow, 2D skeleton, 3D skeleton, and clinician ASD evaluation scores of children. The average length of each video is about 7 s with a 25 to 30 fps frame rate.

LOGO (Zhang, Dai et al., 2023): The LOFO dataset focuses on multi-person long-form artistic swimming scenarios. It contains 200 videos from 26 artistic swimming events with 8 athletes in each sample along with an average length of 204 s. Formation labels to depict group information of multiple athletes, detailed annotations on action procedures and scores are collected.

3.4. Methods

In the last decade, various methods have been proposed for assessing action quality. Generally speaking, based on data modalities, there are two directions to evaluate action quality: skeleton-based methods and video-based methods.

3.4.1. Skeleton-based methods

As skeleton data contains intuitive structural body pose and joint information, which is suitable for action analysis. Thus, in the earlier stage, many studies attempted to evaluate the performance of actions by hand-crafted methods, which calculate the similarity between the skeleton sequence of reference and subjects such as Dynamic Time Warping (DTW) and Euclidean distance. Subsequently, with the advancement of machine learning algorithms, approach such as Support Vector Machines (SVM), decision tree, ridge regression, support vector regression and random forest have been employed for AQA tasks based on skeleton features. In recent years, with the prosperity of deep learning models, Graph Convolutional Networks (GCNs) have been proposed and successfully applied to human AQA based on skeleton topological structures. Meanwhile, the advent of low-cost 3D motion cameras (such as Kinect)and pose-tracking estimators (such as Open Pose) makes acquiring skeleton information much easier, which promotes the development of skeleton-based methods. In this section, we grouped skeleton-based methods into traditional methods and GCN methods.

Traditional methods. Çeliktutan et al. (2013) artificially added different levels of Gaussian noise to the joint positions of graph-based standard performed sequence to acquire the wrong sequence and then utilized SVM to classify correct and wrong sequence for action quality assessment. Venkataraman and Turaga (2016), Venkataraman et al. (2013) designed a Home-based Adaptive Mixed Reality Rehabilitation (HAMRR) system to track wrist movements during therapy treatment. They also trained SVM regression model for evaluating the level of functional ability of stroke patients by using dynamical shape features built from HAMRR system data as input and Functional Activity Score (FAS) from the Wolf Motor Function Test (WMFT) as ground truth. The Pearson correlation coefficient was used to evaluate the performance of proposed AQA framework. In addition, based on the kinematic feature of stroke patients' wrist trajectory, they proposed a hierarchical model using decision trees (Breiman, Friedman, Olshen, & Stone, 1984) to evaluate movement quality of reaching and grasping a cone target (Venkataraman et al., 2014). Furthermore, a linear model was proposed for movement quality assessment based on the kinematic feature of wrist trajectories like trajectory error, jerkiness, velocity and peak speed (Venkataraman et al., 2016). Jin et al. (2016) used Dynamic Space-Time Warping (DSTW) (Yao & Zhu, 2009) and 3D Euclidean space to compute the similarity of joints between user actions and standard actions for action quality assessment. Lei et al. (2020) employed OpenPose to extract skeleton data from RGB videos and then used support vector regression and ridge regression for AQA based on the feature of joint trajectories and joint displacement sequences. Wang, Jin et al. (2020). employed PosePrior to estimate the 3D hand coordinates and utilized Long Short-Term Memory (LSTM), Discrete Cosine Transform with Support Vector Classifier (DCT+SVC), and Discrete Fourier Transform with Support Vector Classifier (DFT+SVC) for human hand action quality assessment. Dajime et al. (2020) used multiclass logistic regression to classify movement quality by domain knowledge-based Kinematic features from Kinect joint position data. Based on Iterative Closest Point (ICP) algorithm (Besl & McKay, 1992; Li, Cui, Guo, Hu, & Shen, 2020), Li, Hu et al. (2021) proposed a local–global geometrical registration strategy by aligning the skeleton sequence from coach and subjects to find difference for fitness action assessment. To achieve pull-ups action quality assessment, Liu, Wang et al. (2023) divided the pull-ups cycle into five distinct states and defined corresponding scoring criteria for each state, and then proposed PEPoseNet to extract human pose features and applied random forest classifier to obtain final score. Hipiny et al. (2023) collected a ranked TikTok dataset and used a pairwise method to predict the better dancer in a pair of videos. Fang, Luo et al. (2023) developed a mixed reality-based game for post-stroke rehabilitation and set some thresholds to calculate the movement quality. They evaluated the proposed method by calculating the Intra-group Correlation Coefficient (ICC) between predicted scores and the ground truth scores from therapists. Wang et al. (2023) fused the DTW and classification results to evaluate the performance of BaDuanJin movements.

GCN. Yan, Xiong, and Lin (2018) first applied GCN for modelling dynamic skeletons and action recognition. They built a spatial–temporal graph model and proposed Spatial–Temporal Graph Convolutional Networks (ST-GCN) based on a sequence of skeleton graphs, where the joints of the human body were considered as nodes, the natural connections between joints as spatial edges, and the connections between the same joint across consecutive frames as temporal edges.

With the outstanding performance of GCN in human action recognition (Yan et al., 2018; Yu et al., 2020) attempted to utilize GCN to detect abnormal actions. They employed GCN based on skeleton data to classify correct and incorrect actions and conducted experiments on UI-PRMD public dataset. In addition, aiming to utilize GCN to monitor the outcomes of behavioural therapies for Alzheimer's disease, Yu, Liu et al. (2021) proposed a tow-task GCN (2T-GCN) based on skeleton data to detect abnormality and predict numerical evaluation scores to indicate the severity of Alzheimer's disease patients. Numerical evaluation scores were calculated from the probability distribution of the final SoftMax layer of 2T-GCN model. They further conducted experiments on both public UI-PRMD dataset and their own Elderly Home Exercise (EHE) dataset to validate the consistency of the predicted score with the clinical evaluation. Li, Hu et al. (2022) proposed a home-based fitness action analysis system with Kinect Azure camera. They collected the RGB-D images and 3D skeletons of 11 subjects performing the 24-form Tai Chi and applied ST-GCN framework to classify different levels of 24-form Tai Chi Quan. To assess action quality in long-term sports videos, Li, Lei et al. (2021, 2022) proposed a spatial–temporal pose feature learning framework for AQA in figure skating videos based on skeleton information estimated by Openpose. First, they designed a Spatial-Temporal Posed Extraction module (STPE) by utilizing ST-GCN (Yan et al., 2018) as the backbone to extract spatial and temporal features of skeletal data. Then, to capture temporal information between skeletal subsequences in long-term videos, an inter-Action Temporal Relation Extraction model (ATRE) implemented by LSTM (Hochreiter & Schmidhuber, 1997) was proposed. Finally, they used a full connect network to regress the final score and evaluated the effectiveness on the MIT-Skate and FIS-V datasets. Zhang, Wang et al. (2023) proposed a Structural-Feature Adaptive Fusion Graph Convolutional Network (SFAGCN) consisting of GCN and TCN blocks for extracting spatio-temporal features, and then employed attention

Table 7
Overview of feature extraction methods.

Method	Numbers	Related work
2D-CNN	10	Baby et al. (2022), Li, Zhang, Dong et al. (2022), Li, Zhang et al. (2023), Li, Zhang, Lei et al. (2022), Matsuyama et al. (2023), Nekoui et al. (2021), Parmar et al. (2021), Wang, Du et al. (2020), Yuan (2024) and Zeng et al. (2020)
I3D	28	Bai et al. (2022), Dadashzadeh et al. (2024), Fang, Zhou et al. (2023), Freire-Obregón et al. (2022), Gao, Pan et al. (2023), Gao et al. (2020), Gedamu et al. (2023), Huang and Li (2024), Ke et al. (2024), Lei et al. (2021), Li, Bhat et al. (2021), Liu, Cheng et al. (2023), Liu, Zhai et al. (2023), Nekoui et al. (2021), Pan et al. (2019, 2022), Roditakis et al. (2021), Tang et al. (2020), Wang, Yang, Zhai, Chen et al. (2021), Xu, Rao et al. (2022), Yu, Rao et al. (2021), Zeng et al. (2020), Zeng and Zheng (2024), Zhang, Pan et al. (2022, 2024), Zhang, Xiong et al. (2022), Zhou, Cai et al. (2023) and Zhou, Ma et al. (2023)
C3D	7	Li et al. (2018), Li, Chai et al. (2019), Li, Cui et al. (2022), Nagai et al. (2021), Parmar and Morris (2019a, 2019b) and Parmar and Tran Morris (2017)
VST	5	Du et al. (2023), Ji et al. (2023), Xu, Zeng et al. (2022), Zeng and Zheng (2024) and Zhang, Dai et al. (2023)
P3D	4	Dong et al. (2021), Xiang et al. (2018) and Zhang, Dong et al. (2022, 2023)
X3D	2	Freire-Obregon et al. (2023) and He et al. (2024)
Attention	13	Bai et al. (2022), Gao, Pan et al. (2023), Ji et al. (2023), Lei et al. (2021), Li, Cui et al. (2022), Liu, Cheng et al. (2023), Nekoui et al. (2021), Sun, Hu et al. (2023), Wang, Du et al. (2020), Wang, Yang, Zhai, Chen et al. (2021), Xu, Zeng et al. (2022), Zeng et al. (2020) and Zhang, Pan et al. (2024)

mechanism to fuse spatial features extracted by GCN, temporal features extracted by TCN and spatio-temporal features extracted by SFAGCN for action quality assessment. Joung et al. (2023) proposed a virtual joint-based GCN to learn motion information based on contrast learning. In addition, a Degraded Negative Contrasting method was proposed to improve the model performance in contrast learning. To leverage pose feature, Lei et al. (2023) proposed a multi-skeleton structures graph convolutional network (MS-GCN) and combined with a temporal attention block for long-duration activity. Li, Ling et al. (2023) combined multi-task learning and contrastive learning. They proposed a GCN-based siamese network to predict action recognition and action assessment simultaneously from a pair of exercises. Based on GCN, Mourchid and Slama (2023) used multiple residual layers to extract features from different levels and intergrated the attention mechanism for feature enhancement. Zhang, Wang et al. (2023) proposed structure-feature fusion adaptive graph convolutional networks (SFAGCN) which employed the attention mechanism to adaptive fuse spatial and temporal features. Yu et al. (2024) proposed an ensemble-based graph convolutional network and explored the fusion strategy in graph-based models for skeleton-based exercise assessment. They employed GCN-based methods to assess exercises and fused information in different levels including data, model, feature and decision levels.

In addition to GCN, some studies applied other deep learning methods to extract skeleton features. Kanade et al. (2023a) utilized transformer encoder after the CNN-based extractor to improve the performance in physical exercise assessment. Furthermore, they also combined CNN and LSTM for AQA in the following work (Kanade et al., 2023b). Chariar et al. (2023) collected a dataset and proposed a network based on LSTM and attention for squat analysing. Li, Tian et al. (2023) applied transformer for martial arts recognition and evaluation. Hirosawa et al. (2023) introduced the specialist's gaze information and employed VGG for predicting jump execution scores in figure skating. Zhang, Zhou et al. (2023) combined CNN and LSTM methods to predict the Autism Diagnostic Observation Schedule (ADOS) score based on skeletal data for autism spectrum disorder analysis. Sardari et al. (2024) proposed a Light Physical Rehabilitation Assessment (LightPRA) method based on TCN to improve computational efficiency.

3.4.2. *Video-based methods*

The development of feature extraction methods in the video-based AQA is closely related to video understanding, which can be generally categorized into 2D-CNN, 3D-CNN and attention mechanisms. In the early stages, inspired by the breakthrough of CNN in the image domain (Krizhevsky, Sutskever, & Hinton, 2012; LeCun, Bengio, et al., 1995), numerous studies have begun applying CNNs to extract features for video understanding (Gkioxari & Malik, 2015; Hou, Chen, & Shah, 2017; Jain, Tompson, LeCun, & Bregler, 2015; Karpathy et al., 2014; Peng & Schmid, 2016; Zolfaghari, Oliveira, Sedaghat, & Brox, 2017), thereby advancing the development of video-based AQA methods. Some 2D-CNNs such as ResNet, TCN, SCN are commonly used to extract spatial and temporal features in frames for AQA. Then, with the advent of 3D networks, Convolution 3D networks (C3D) (Tran, Bourdev, Fergus, Torresani, & Paluri, 2015), Inflated 3D convolutional networks (I3D) (Carreira & Zisserman, 2017) and Pseudo-3D ResNet (P3D) (Qiu, Yao, & Mei, 2017), are commonly used to capture spatiotemporal features from RGB videos and then complete a regression or classification task for AQA. Currently, with the advanced ability of transformer and attention mechanisms in capturing global temporal relationships, some studies have applied attention mechanisms in AQA. In Table 7, we summarized the commonly used feature extraction methods in video-based AQA based on included 96 papers. We can find I3D is the most commonly used feature extraction method in AQA. It is worth noting that in recent years, there has been a growing trend in studies that employ attention mechanisms to enhance feature extraction capability.

As we mentioned before, AQA needs to analyse the whole video and capture the internal fine-granularity feature of an action, which is more challenging. Thus, a lot of frameworks have been proposed for different AQA tasks.

Some studies proposed to evaluate the action performance stage by stage as each stage contributes differently to the overall score. For example, a complete diving performance can be divided into five stages: beginning, jumping, dropping, entering into water and ending, and each stage contributes differently to the total score. Xiang et al. (2018) Firstly rated diving actions stage by using Encoder–Decoder Temporal Convolutional Network (ED-TCN) (Lea, Flynn, Vidal, Reiter, & Hager, 2017; Lea, Vidal, Reiter, & Hager, 2016) to generate 5 segments of full videos, using 4 P3D models to extract feature presentations of 4 crucial segments and then using fully connected layer, SVR or LR to predict the score based on average of all stage-wise features. Dong et al. (2021) proposed a multiple-substage network for diving action quality assessment. Firstly, they applied ED-TCN to segment videos into five stages and employed P3D residual network to extract spatial and temporal features of each stage, and then fed into five decreasing FC networks to generate five substage score features. Finally, a liner regression model was used to predict the final score. In addition, Zhang, Dong et al. (2022) proposed a label-reconstruction-based Pseudo Subscore Learning (PSL) method to build pseudo-substage sore by setting the activation function of the last layer in FC networks as Sigmoid function. By dividing diving into multiple phrases and summing the predicted scores of individual segments, Liu, Zhai et al. (2023) improved the perfomance on baseline AQA models. To improve the interpretability of AQA model, Matsuyama et al. (2023) proposed an interpretable rubric-informed segmentation (IRIS) network to segment figure skating into different category sequences and predict the score in figure skating. Huang and Li (2024) proposed Semantic-Sequence Performance Regression network to enhance the semantic information by segmenting an action into unequal-length clips. Zhang, Pan et al. (2024) proposed an adaptive stage-aware assessment skill transfer (AdaST) framework to transfer assessment knowledge among relevant skills.

Some studies applied multi-task learning framework in AQA, Parmar and Morris (2019b) utilized C3D model to realize action quality score, action recognition and commentary tasks simultaneously and they found multitask learning can achieve better performance in AQA compared with Single AQA task because of better generalization. Based on the multi-task learning framework, Zhang, Xiong et al. (2022) combined AQA task with caption generation auxiliary task and action recognition auxiliary task to obtain better generalization. In addition, they utilized I3D to extract features of video clips and proposed to add adversarial loss into Time-aware attention mechanism to capture the relationship between different video clips for AQA task. Li, Bhat et al. (2021) utilized I3D to extract features from child and therapist video pairs and employed multi-task learning to realize action quality assessment task, movement synchrony estimation task and intervention action recognition task simultaneously. Gan et al. (2024) proposed a large-scale benchmark SkatingVerse that covers multiple tasks in action understanding including action localization, recognition and assessment.

Some studies applied contrastive regression framework in AQA. To capture the subtle difference, Yu, Rao et al. (2021) proposed a Contrastive Regression (CoRe) framework based on residual learning for AQA. They utilized I3D to extract features of the input videos and exemplar videos and aggregate them with the score of exemplar videos, and then employed group-aware regression tree to regress the difference of the scores between the input videos and exemplar videos. Based on Yu's work, Bai et al. (2022) proposed a Temporal Parsing Transformer (TPT) to extract fine-grained temporal part-level features and combined them with contrastive regression framework for AQA. In addition, Li, Zhang, Lei et al. (2022) employed ResNet and a temporal encoder network to extract features and combined with a new proposed Pairwise Contrastive Learning Network (PCLN) for AQA. And then, to improve the reliability and obtain better interpretability, Xu, Rao et al. (2022) proposed a procedure-aware approach by utilizing transformer decoder (Dosovitskiy et al., 2020) based on the contrastive regression framework for AQA and demonstrated its effectiveness on FineDiving the first fine-grained sports video dataset constructed by them. Liu, Cheng et al. (2023) proposed a triple-stream contrastive transformer which introduces a replay branch to learn features under different views. Zhou, Cai et al. (2023). proposed a contrastive-based AQA method and applied it to the proposed video-based augmented reality visualization system for juvenile dermatomyositis assessment. Ke et al. (2024) proposed a two-path target-aware contrastive regression framework by fusing direct loss and contrastive loss to improve assessment performance.

In addition, some studies applied distribution learning framework in AQA. Inspired by the fact that the final score in diving is uncertain and obtained by multiple judges, Tang et al. (2020) proposed an Uncertainty-Aware score Distribution Learning (USDL) method for AQA. They utilized I3D to extract features from video clips and learn the Gaussian distribution of scores rather than a single final score label, to directly predict the score distribution of an action. Furthermore, considering the final score was calculated based on the execution score of each judge and the difficulty degree, a multi-path uncertainty-aware score distributions learning (MUSDL) method was proposed to leverage these components. Lei et al. (2021) used I3D to extract features of evenly divided video segments and then employed the attention mechanism and distribution learning strategy of Tang et al. (2020) to learn temporal weights on different action stages to balance significance of different segments in sports video for action quality assessment. Inspired by Tang, Li, Zhang, Dong et al. (2022) applied Gaussian to model the score label instead of MSE and utilized ResNet and a frame sequence-based temporal encoder convolutional network to extract temporal and spatial features of full-video frames instead of video clips for AQA. Ji et al. (2023) proposed a Localization-assisted Uncertainty Score Disentanglement Network (LUSD-Net) that locates technical subaction and utilizes uncertainty regression to enhance feature disentanglement for scoring figure skating. Li, Zhang et al. (2023) used the Gaussian loss function to compute the error between the predicted score and the label score. Lian and Shao (2023) employed kernel density estimation to reweight labels to solve the data imbalanced issue in regression for action quality assessment. Zhang, Chen et al. (2024) proposed a distribution auto-encoder module to learn uncertainty and reduce the imbalance of data.

Some studies focus on modelling asymmetric interaction among agents for AQA as there are asymmetric relations among agents (e.g., between subjects and objects) in non-individual actions in our real-world scenarios. Gao et al. (2020) used attention mechanism to combine whole-scene features extracted by I3D and the asymmetric interactions between agents within an action extracted by proposed Asymmetric Interaction Module (AIM) for interactive action quality assessment. In the subsequent study (Gao, Pan et al., 2023), they improved the Asymmetric interaction module by proposing an automatic assigner and an asymmetric interaction network search module. The former can automatically discriminate primary and secondary agents, while the latter can adaptively learn the asymmetric interactions between these agents. Furthermore, they collected two new datasets, TASD-2 for synchronous diving and PaSk for pair figure skating.

Some studies try to explore self-supervised and semi-supervised methods in AQA as acquiring domain experts' professional annotations is challenging in AQA field. Zhang, Pan et al. (2022) first applied semi-supervised learning and proposed a self-supervised based framework for action quality assessment, which can complete AQA task with only a small amount of labelled data. Similarly, Roditakis et al. (2021) proposed a self-supervised method. They utilized Temporal Cycle Consistency (TCC) learning (Dwibedi, Aytar, Tompson, Sermanet, & Zisserman, 2019) method to realize video alignment and combined TCC and I3D to extract features for action quality assessment. Zhang, Dong et al. (2023) developed a label-reconstruction-based pseudo-subscore learning (PSL) method for the lack of substage quality in AQA. He et al. (2024) introduced a weakly supervised framework to learn rich correlation information between two videos and improved AQA performance.

Some studies attempted to extract features from human motions rather than background scenes as humans occupy only a small part of scenes in videos. Zeng et al. (2020) proposed a hybrid dynAmic-static Context-aware attenTION NETwork (ACTION-NET) to capture both video dynamic information and static posture information for AQA in skating and rhythmic gymnastics long videos. First, they utilized I3D networks to extract dynamic feature from video segments and static posture features from some sampled frames processed by human detection. Moreover, a context-aware attention module consisting of a temporal instance-wise graph convolutional network unit and an attention unit was used to aggregate these features. Finally, a fully connected layer was used to concatenate features and regress the final score. Pan et al. (2019) considered the detailed joint interaction and proposed Joint Relation Graph (JRG) to extract the difference and commonalty features of joints during motion. Based on I3D and regression model, they extract features of the whole scene videos and local joints videos for assessing Olympic events. in addition, considering the judging criteria are different in different sports events such as diving, gymnastics, and snowboarding. Pan et al. (2022) proposed adaptive action assessment approach, by utilizing a differentiable network architecture search mechanism (Liu, Simonyan, & Yang, 2018) to adaptively design different architectures for different actions. Nekoui et al. (2021) utilized a two-stream network (EAGLE-Eye) comprising of proposed Joints Coordination Assessor (JCA) block and proposed Appearance Dynamics Assessor (ADA) blocks to capture both appearance and pose features for AQA. Considering the long-term AQA tasks, They also proved that empowering the network with more JCA and ADA blocks can capture long-term global features. To mitigate the impact of scene information, Nagai et al. (2021) imposed scene adversarial loss and human-masked regression loss to C3D-AVG model for action quality assessment with ignoring scene context.

In order to learn more features, scholars have applied attention mechanism and transformer in AQA. Wang, Du et al. (2020) utilized Spatial Convolutional Network (SCN) (Szegedy, Vanhoucke, Ioffe, Shlens, & Wojna, 2016) and Temporal Convolutional Network (TCN) (Lea et al., 2017) to extract spatio-temporal features and then employed an attention mechanism to fuse features in temporal dimensions for AQA task. Wang, Yang, Zhai, Chen et al. (2021) adopted self-attention mechanism and proposed a Tube Self-Attention Network (TSA-Net) to extract more contextual information from videos for AQA. They utilized the proposed TSA mechanism to aggregate features extracted from I3D and tracking results from Visual Object Tracking (VOT) tracker, and then fed these features into I3D for action quality assessment. Li, Cui et al. (2022) utilized 3D ResNet to extract features and used Multilayer Perception (MLP), Variational Autoencoder (VAE) and Vision Transformer to regress the gymnastic score. Xu, Zeng et al. (2022) proposed a Grade-decoupling Likert Transformer model. They adopted transformer (Liu et al., 2022; Vaswani et al., 2017) encoder–decoder architecture to decouple a long-term action video into different level features which can obtain different corresponding grades, and then applied Likert Scoring (Likert, 1932) to aggregate different grades to generate the final quality score. Sun, Hu et al. (2023) proposed GRU-based spatial and temporal pooling methods to enhance the spatio-temporal features from the feature extractor. Fang, Zhou et al. (2023) employed transformer encoder to enhance the context information between the features extracted by I3D. Gedamu et al. (2023) introduced a Fine-grained spatio-temporal pasing network (FSPN) to extract subtle intra-class variation in AQA. Dadashzadeh et al. (2024) proposed a parameter efficient, continual pretraining (PECoP) framework. They introduce learnable 3D adapters which can learn in-domain information to enhance generalization ability when transferring knowledge. Zhou, Ma et al. (2023) proposed a hierarchical GCN framework dedicated to learning the semantic context across video clips. Zhang, Dai et al. (2023) built a novel multi-person long-form video (LOGO) dataset and proposed a group-aware model to capture relations among multiple athletes.

Some studies fused multiple modalities for AQA. Parmar et al. (2021) try to assess piano skills by concatenating video features and audio features. Wang, Yang, Zhai, and Zhang (2024) proposed a multimodel framework CPR-CLIP to recognize incorrect external cardiac compression action in Cardiopulmonary Resuscitation skill training. Nagai et al. (2024) developed a novel multimodal in-the-wild (MMW)-AQA dataset and proposed a transformer-based baseline model for freestyle windsurfing assessment. In addition to RGB video, Inertial Measurement Unit (IMU) data and Global Positioning System (GPS) data are provided in the dataset. Zeng and Zheng (2024) considered the audio information for sports with background and proposed a Progressive Adaptive Multimodal Fusion Network (PAMFN) that fused RGB,optical flow and Audio information for AQA in figure skating and rhythmic gymnastics.

3.5. Evaluation metrics

Taking into account the difference between the evaluation metrics applied in different types of data, we studied the selected 96 research articles and categorized the evaluation metrics into skeleton-based and video-based metrics.

3.5.1. Skeleton-based metrics

We present the overview of evaluation metrics in skeleton-based methods, as shown in Table 8. We can find there are no uniform metrics to evaluate skeleton-based AQA models, as the definition of evaluation tasks varies in different studies. Some studies define AQA as a regression task such as scoring the movements of athletes in Olympic games. Therefore, the performance of AQA models is evaluated by calculating the correlation between predicted scores from models and ground truth scores. In skeleton-based methods, Spearman's Rank Correlation (SRC), Pearson Correlation Coefficient (PCC), Mean Rank Correlation (MRC) and Intra-group Correlation Coefficient (ICC) are used to calculate the similarity between ground truth and predicted results. Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Euclidean Distance (MED), Mean Absolute Error (MAE) and Euclidean Distance (ED) are used to measure the error between the ground truth and predicted results. Additionally, some studies label the quality of actions into different discrete levels which can be formulated as a classification problem and Classification Accuracy (CA), F1-score, Precision, are employed to evaluate the skeleton-based AQA models. The most commonly used evaluation metric is SRC, which is employed in a total of 13 articles, followed by CA, which is used in 10 articles. The Spearman's rank correlation (SRC) is defined as:

$$\rho = \frac{\sum_i (p_i - \bar{p})(q_i - \bar{q})}{\sqrt{\sum_i (p_i - \bar{p})^2 \sum_i (q_i - \bar{q})^2}} \tag{1}$$

3.5.2. Video-based metrics

In Table 9, we summarized the evaluation metrics in video-based methods. We can observe that Spearman's Rank Correlation (SRC), Pearson Correlation Coefficient (PCC) and Kendall Correlation (KC) are used to calculate the similarity between ground truth and predicted results. Additionally, Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Euclidean Distance (MED), Mean Absolute Error (MAE), and the relative ℓ_2-distance (R-ℓ_2) are used to measure the error between the ground truth and predicted results. Classification Accuracy (CA), mean Average Precision (mAP) and mean Average Precision of the multi-modal image-text model ($mmit$ mAP) are employed to evaluate the classification tasks in video-based AQA models. SRC is the most frequently used evaluation metric, employed in a total of 56 articles, which is 45 more than the second most used metric, R-ℓ_2. The relative ℓ_2-distance (R-ℓ_2) is defined as:

$$R\text{-}\ell_2 = \frac{1}{N}\sum_{n=1}^{N}\left(\frac{|s_n - \hat{s}_n|}{s_{max} - s_{min}}\right)^2 \tag{2}$$

Due to the fact that SRC is the most commonly used evaluation metric in both skeleton-based and video-based methods, we present the SRC performance of some state-of-the-art methods on three commonly used datasets: MTL-AQA, AQA-7 and JIGSAW datasets, which are shown in Tables 10–12. It is worth noting that in MTL-AQA diving dataset, studies in Bai et al. (2022), Tang et al. (2020) and Yu, Rao et al. (2021) do not directly predict the final score. Instead, they predict the execution score first and then multiply it by the difficulty degree label to obtain the final score, resulting in two sets of results (one with difficulty degree labels and another without). For AQA-7 dataset, the Fisher's z-value (Faller, 1981) is used to measure the average SRC across actions.

Table 8
Overview of evaluation metrics in skeleton-based methods.

Evaluation metrics	Numbers	Related paper
SRC	13	Gao, Pan et al. (2023), Hirosawa et al. (2023), Lei et al. (2023), Li, Hu et al. (2021), Li, Lei et al. (2021, 2022), Nekoui et al. (2021), Pan et al. (2019), Wang et al. (2022), Yu et al. (2024), Yu, Liu et al. (2021), Zhang, Wang et al. (2023) and Zhang, Zhou et al. (2023)
CA	10	Çeliktutan et al. (2013), Chariar et al. (2023), Dajime et al. (2020), Li, Hu et al. (2022), Li, Ling et al. (2023), Li, Tian et al. (2023), Liu, Wang et al. (2023), Wang, Jin et al. (2020), Wang et al. (2022) and Yu et al. (2020)
PCC	5	Venkataraman and Turaga (2016), Venkataraman et al. (2016, 2013, 2014) and Wang et al. (2022)
MAE	4	Hirosawa et al. (2023), Kanade et al. (2023a), Mourchid and Slama (2023) and Sardari et al. (2024)
RMSE	3	Hirosawa et al. (2023), Kanade et al. (2023b) and Sardari et al. (2024)
ED	3	Jin et al. (2016), Yu et al. (2024) and Yu, Liu et al. (2021)
MRC	2	Lei et al. (2020) and Pirsiavash et al. (2014)
MSE	1	Li, Lei et al. (2022)
MED	1	Li, Lei et al. (2021)
ICC	1	Fang, Luo et al. (2023)
F1-score	1	Joung et al. (2023)
Precision	1	Hipiny et al. (2023)

Table 9
Overview of evaluation metrics in video-based methods.

Evaluation metrics	Numbers	Related paper
SRC	56	Baby et al. (2022), Bai et al. (2022), Dadashzadeh et al. (2024), Dong et al. (2021), Du et al. (2023), Fang, Zhou et al. (2023), Farabi et al. (2022), Gan et al. (2024), Gao, Pan et al. (2023), Gao et al. (2020), Gedamu et al. (2023), He et al. (2024), Hirosawa et al. (2023), Huang and Li (2024), Ji et al. (2023), Ke et al. (2024), Lei et al. (2021), Li, Bhat et al. (2021), Li et al. (2018), Li, Chai et al. (2019), Li, Zhang, Dong et al. (2022), Li, Zhang et al. (2023), Li, Zhang, Lei et al. (2022), Lian and Shao (2023), Liu, Cheng et al. (2023), Liu, Zhai et al. (2023), Matsuyama et al. (2023), Nagai et al. (2021, 2024), Nekoui et al. (2021), Pan et al. (2019, 2022), Parmar and Morris (2019a, 2019b), Parmar and Tran Morris (2017), Roditakis et al. (2021), Sun, Hu et al. (2023), Tang et al. (2020), Wang, Du et al. (2020), Wang, Yang, Zhai, Chen et al. (2021), Xiang et al. (2018), Xu, Rao et al. (2022), Xu, Zeng et al. (2022), Yu, Rao et al. (2021), Zeng et al. (2020), Zeng and Zheng (2024), Zhang, Chen et al. (2024), Zhang, Dai et al. (2023), Zhang, Dong et al. (2022, 2023), Zhang, Pan et al. (2022, 2024), Zhang, Wang et al. (2023), Zhang, Xiong et al. (2022), Zhou, Cai et al. (2023) and Zhou, Ma et al. (2023)
R-ℓ_2	11	Bai et al. (2022), Fang, Zhou et al. (2023), Gedamu et al. (2023), He et al. (2024), Ke et al. (2024), Lian and Shao (2023), Liu, Cheng et al. (2023), Xu, Rao et al. (2022), Zhang, Dai et al. (2023), Zhou, Cai et al. (2023) and Zhou, Ma et al. (2023)
MSE	7	Du et al. (2023), Ingwersen et al. (2023), Li, Bhat et al. (2021), Li, Cui et al. (2022), Yuan (2024) and Zhang, Dong et al. (2022, 2023)
MED	5	Lei et al. (2021), Li et al. (2018), Li, Chai et al. (2019) and Zhang, Dong et al. (2022, 2023)
CA	3	Freire-Obregón et al. (2022), Parmar et al. (2021) and Zhang, Pan et al. (2024)
MAE	3	Freire-Obregon et al. (2023), Hirosawa et al. (2023) and Sun, Hu et al. (2023)
PCC	1	Matsuyama et al. (2023)
RMSE	1	Hirosawa et al. (2023)
KC	1	Ingwersen et al. (2023)
mAP and $mmit$ mAP	1	Wang et al. (2024)

4. Discussion

This section analyses the research results and presents findings according to the research questions defined in Section 2. In addition, the limitation of this review is discussed.

4.1. Summary of reviewed studies

To present a systematic literature review on vision-based human AQA, we defined 5 research questions related to the application, data acquisition, datasets, methods and evaluation metrics in Section 2. 96 research articles were studied to answer the research questions. The main findings are as follows:

4.1.1. Application

Table 2 presents the existing applications of AQA and related studies, which can enable researchers to efficiently find research related to their specific field of interest. As illustrated, AQA has been applied to various scenarios but mainly focuses on three scenarios: sports event scoring, surgical skill evaluation and rehabilitation assessment. Notably, sports event scoring has attracted the most attention, with 53 out of the 96 selected papers dedicated to this field. This is because there are some publicly available datasets in the sports scoring domain, which can facilitate research in this area. Due to the growing demand for movement assessment, AQA is now being applied to an increasing number of new domains, such as behaviour therapy, fitness, exercise, piano, hand skill, daily action, windsurfing and dance etc. In these applications, AQA can reduce the workload of professionals, explore new supervision and feedback mechanisms, and enhance the efficiency of training. Specifically, in medical rehabilitation, AQA can provide precise assessments of patients' progress, allowing therapists to tailor treatment plans more effectively. In the fitness industry, AQA can offer real-time feedback to improve workout effectiveness and reduce the risk of injuries. The application of AQA in new domains like dance, windsurfing, and piano demonstrates its versatility and adaptability.

4.1.2. Data acquisition

We can find optical motion capture systems, depth cameras and pose estimation algorithms are commonly used methods to obtain skeleton data in Table 3. Specifically, pose estimation algorithms present the

Table 10
Overview of the performance of state-of-the-art methods in MTL-AQA dataset. The bold number indicates the best performance. 'w/o DD' means without difficulty degree labels, 'w/o DD' means with difficulty degree labels.

Methods (w/o DD)	Year	SRC
Pose+DCT (Pirsiavash et al., 2014)	2014	0.2682
C3D-SVR (Parmar & Tran Morris, 2017)	2017	0.7716
C3D-LSTM (Parmar & Tran Morris, 2017)	2017	0.8489
MSCADC-STL (Parmar & Morris, 2019b)	2019	0.8472
MSCADC-MTL (Parmar & Morris, 2019b)	2019	0.8612
C3D-AVG-STL (Parmar & Morris, 2019b)	2019	0.8960
C3D-AVG-MTL (Parmar & Morris, 2019b)	2019	0.9044
USDL (Tang et al., 2020)	2020	0.9066
CoRe (Yu, Rao et al., 2021)	2021	0.9341
TSA-NeT (Wang, Yang, Zhai, Chen et al., 2021)	2021	0.9422
TPT (Bai et al., 2022)	2022	0.9451
HGCN (Zhou, Ma et al., 2023)	2023	0.9390
T^2CR (Ke et al., 2024)	2024	**0.9464**
Methods (w/DD)	**Year**	**SRC**
USDL (Tang et al., 2020)	2020	0.9231
MUSDL (Tang et al., 2020)	2020	0.9273
CoRe (Yu, Rao et al., 2021)	2021	0.9512
TPT (Bai et al., 2022)	2022	0.9607
HGCN (Zhou, Ma et al., 2023)	2023	0.9536
T^2CR (Ke et al., 2024)	2024	**0.9638**

most used approach to obtain skeleton data. This is because pose estimation algorithms can capture joint coordinates through standard RGB cameras in real time, reducing the need for expensive hardware. Depth cameras are also popular for capturing skeleton data because of their low cost and flexibility. However, the optical motion capture system is the least option for capturing. Although it can obtain the highest tracking accuracy, it also requires specialized equipment and controlled environments, which limits its application. In conclusion, the emergence of commercial depth cameras and pose estimation algorithms enables us to obtain skeleton information easily for AQA tasks, which can promote the development of datasets and the application of AQA.

4.1.3. Datasets

As shown in Table 6, we can observe many datasets in various domains, including sports scoring, rehabilitation and exercise evaluation, surgical skill rating, piano skill rating, etc., have been proposed in recent years, indicating the increasing attention to AQA research. However, most of them focus on Olympic event scoring, while a few focus on medical care and other action categories. The reason is probably that recruiting domain experts to annotate data can also be time-consuming and expensive (Roditakis et al., 2021). The Olympics event will publicize game videos and provide the ground truth of the judges' scores, allowing researchers to build datasets more easily. In addition, privacy concern is a challenge in constructing medical AQA datasets. In the early stages, AQA datasets typically provided only an overall score in their annotations, which limited the interpretability of the models. The recently proposed AQA datasets (Ji et al., 2023; Xu, Rao et al., 2022; Zhang, Dai et al., 2023) provide fine-grained annotations to enhance the interpretability of AQA models. This increased level of detail allows for a more nuanced understanding of the assessed actions, giving researchers richer information for model training and evaluation. Furthermore, existing AQA datasets primarily focus on single-person with short-duration actions, with comparatively fewer datasets addressing multi-person with long-duration actions. Additionally, most datasets provide either video or skeletal data exclusively, with only a few datasets offering multiple types of data.

4.1.4. Methods

The AQA methods can be divided into skeleton-based methods and video-based methods.

For skeleton-based methods, handcrafted methods were applied for AQA based on skeleton data collected by professional motion tracking systems in the earlier stage (Çeliktutan et al., 2013; Venkataraman et al., 2013, 2014). These methods can be able to solve two binary classification AQA problems (e.g., correct/incorrect), but it is difficult to design discriminate rules for complex activities in long-duration videos. Due to the rapid development of deep learning and the advent of low-cost 3D motion cameras (e.g., Kinect), as well as pose-tracking estimators (e.g., OpenPose, MediaPipe), GCN has been successfully applied to human AQA based on the extracted skeleton data and attracted more attention (Li, Lei et al., 2021, 2022; Yu et al., 2020; Yu, Liu et al., 2021). These methods can evaluate human action quality by only focusing on postures and ignoring background information.

For video-based methods, as shown in Table 7, some general network backbones including 2D-CNN, I3D, C3D, P3D and attention mechanisms have been applied for capturing spatiotemporal features from RGB videos in AQA (Gao et al., 2020; Parmar & Tran Morris, 2017; Xiang et al., 2018; Xu, Zeng et al., 2022). I3D is the most commonly used feature extraction method. Furthermore, various strategies and frameworks have been proposed for AQA tasks, such as stage-by-stage strategy, multi-task learning framework, contrastive regression framework, distribution learning framework, asymmetric interaction framework, self-supervised based framework, motion-focused strategy, attention-based strategy and multi-modal models. These methods can assess complicated movement and predict a specific score by regressing extracted features from the whole video.

4.1.5. Evaluation metrics

We found Skeleton-based algorithms and video-based algorithms share both similarities and differences in their choice of evaluation metrics. We found that different metrics can be used to evaluate the performance of AQA models, depending on the definition of action quality in both skeleton-based and video-based methods. Spearman rank correlation (SRC) coefficient is the most commonly used metric in both skeleton-based and video-based methods for calculating the correlation between predicted scores and ground truth scores (Li, Chai et al., 2019; Pan et al., 2019; Parmar & Morris, 2019a). Classification accuracy is commonly used to evaluate AQA models that predict action quality at discrete levels (Çeliktutan et al., 2013; Dajime et al., 2020; Yu et al., 2020). However, R-ℓ_2 is currently the second most frequently used metric in video-based methods but has not yet been employed in skeleton-based methods. AQA has been defined as either a regression or classification task in both methods. However, the proportion of AQA models defined as classification tasks is higher in skeleton-based methods compared to video-based methods.

4.2. Potential threats to validity and limitations

Our systematic literature review only includes 96 of 230 research articles retrieved from three main STEM databases: IEEE Explore, Web of Science and Scopus. However, some studies published only in other databases and some non-English studies were not considered in this paper, which may impact us to provide more insight into AQA.

We provide the performance of some state-of-the-art methods on three commonly used datasets: MTL-AQA, AQA-7 and JIGSAW, due to a sufficient number of studies have applied the same metric to these datasets. However, the comparison of different models in additional datasets is not shown in this paper due to the available space.

5. Current challenges and future research

In Fig. 2, We can observe there is an increase in the number of studies in AQA since 2019. This can be concluded for these reasons: (1) The emergence of commercial depth cameras and pose estimation algorithms enables us to obtain skeleton information easily for AQA

Table 11
Overview of the performance of state-of-the-art methods in AQA-7 dataset. "–" means that the result did not provide in the literature. Bold numbers are the best performances.

Methods	Year	Diving	Gym vault	Skiing	Snow board	Sync. 3 m	Sync. 10 m	Avg. SRC
Pose+DCT (Pirsiavash et al., 2014)	2014	0.5300	–	–	–	–	–	–
ST-GCN (Yan et al., 2018)	2018	0.3286	0.5770	0.1681	0.1234	0.6600	0.6483	0.4433
C3D-LSTM (Parmar & Tran Morris, 2017)	2019	0.6047	0.5636	0.4593	0.5029	0.7912	0.6927	0.6165
C3D-SVR (Parmar & Tran Morris, 2017)	2019	0.7902	0.6824	0.5209	0.4006	0.5937	0.9120	0.6937
JRG (Pan et al., 2019)	2019	0.7630	0.7358	0.6006	0.5405	0.9013	0.9254	0.7849
USDL (Tang et al., 2020)	2020	0.8099	0.7570	0.6538	**0.7109**	0.9166	0.8878	0.8102
EAGLE-Eye (Nekoui et al., 2021)	2021	0.8331	0.7411	0.6635	0.6447	0.9143	0.9158	0.8140
CoRe (Yu, Rao et al., 2021)	2021	0.8824	0.7746	0.7115	0.6624	0.9442	0.9078	0.8401
TSA-Net (Wang, Yang, Zhai, Chen et al., 2021)	2021	0.8379	0.8004	0.6657	0.6962	0.9493	0.9334	0.8476
Adaptive (Pan et al., 2022)	2021	08 306	0.7593	0.7208	0.6940	0.9588	0.9298	0.8500
TPT (Bai et al., 2022)	2022	**0.8969**	0.8043	0.7336	0.6965	0.9456	0.9545	0.8715
PCLN (Li, Zhang, Lei et al., 2022)	2022	0.8697	**0.8759**	**0.7754**	0.5778	**0.9629**	0.9541	**0.8795**
HGCN (Zhou, Ma et al., 2023)	2023	0.8867	0.7917	0.7326	0.6447	0.9213	0.9424	0.8501
T^{2}CR (Ke et al., 2024)	2024	0.8901	0.8393	0.7139	0.7052	0.9418	**0.9558**	0.8726

Table 12
Overview of the performance of state-of-the-art methods in JIGSAW dataset. Bold numbers are the best performances.

Methods	Year	S	NP	KT	Avg. SRC
ST-GCN (Yan et al., 2018)	2018	0.31	0.39	0.58	0.43
JRG (Pan et al., 2019)	2019	0.36	0.54	0.75	0.57
USDL (Tang et al., 2020)	2020	0.64	0.63	0.61	0.63
MUSDL (Tang et al., 2020)	2020	0.71	0.69	0.71	0.70
AIM (Gao et al., 2020)	2020	0.63	0.65	0.82	0.71
CoRe (Yu, Rao et al., 2021)	2021	0.84	0.86	0.86	0.85
TPT (Bai et al., 2022)	2022	0.88	0.88	**0.91**	0.89
HGCN (Zhou, Ma et al., 2023)	2023	0.89	**0.91**	0.90	0.90
T^{2}CR (Ke et al., 2024)	2024	**0.93**	0.89	0.89	**0.91**

tasks. (2) The advent of public datasets encourages more data-driven-based research. (3) The development of deep learning methods, such as 3D networks and attention mechanisms, makes it possible to capture the internal fine-granularity feature of actions in the RGB video for assessment of actions. Although there has been significant progress over the past few years, various challenges still exist in developing vision-based models to automatically assess action quality in real scenarios. We will explain below.

5.1. Application

For many real-world applications, especially in sports and fitness, real-time processing is crucial. However, most existing video-based methods require huge computing costs and time, limiting their practical application. Additionally, the application of AQA in some common activities, such as dancing, football and basketball remains quite limited and requires more in-depth exploration. Conducting additional research in these areas could unveil new possibilities and enhance the effectiveness of AQA in various real-world scenarios.

In the future, AQA applications can be integrated with virtual Reality (VR) and Augmented Reality (AR) technologies to provide richer, more interactive experiences for scenarios requiring movement quality assessment. With the advent of generative models (Guo et al., 2024; Hong, Ding, Zheng, Liu, & Tang, 2022), research on generating actions is emerging, and AQA can be employed to evaluate the quality of these generated actions, thereby enhancing the performance of AI generative models.

5.2. Dataset

Regarding the public available AQA datasets shown in Table 6, there are several challenges. First, most existing AQA datasets are limited to small size and a single type of action, as the AQA criteria vary for different types of actions and the annotations of AQA datasets are typically conducted by domain experts, which can be time-consuming and expensive (Roditakis et al., 2021; Zhang, Pan et al., 2022). Thus, building a diverse and large-scale dataset is still a challenge in AQA. There are some similarities in the evaluation of different actions and skills, which can be useful in training models. In the future, we can create a large dataset encompassing a broader range of action types and a larger number of samples. This can benefit the development of larger models with enhanced generalization capabilities. Second, most existing AQA datasets only provide RGB video information, which is limited to the data modality. In addition to RGB video information, other modalities such as depth information, skeletal data, audio, text and wearable sensor data can also be utilized for assessing action quality. With the development of multi-modal learning, such as Contrastive Language–Image Pre-training (CLIP) (Radford et al., 2021) and Align and Prompt (ALPRO) (Li, Li, Li, Niebles and Hoi, 2022), future datasets can include more data modalities to facilitate future research. The combination of different types of data enables a more comprehensive analysis by capturing the intricacies of both appearance and movement dynamics. Third, the majority of existing datasets offer limited annotations, typically including only an overall score, which constrains the interpretability of the models. In future AQA dataset construction, incorporating more detailed annotations such as sub-action types, temporal boundaries, and sub-scores can further enhance model interpretability. Fourth, datasets specifically designed for the assessment of multi-person actions, such as group dances, are limited and deserve further research. Fifth, current datasets are mostly derived from real-world data. With the advancement of generative models, there is a notable lack of datasets that assess generated actions. In the future, integrating generated data with real-world data could expand both the size and diversity of AQA datasets, providing a more comprehensive resource for model development and evaluation.

5.3. Methods

Although existing AQA methods have achieved remarkable progress, many efforts are still required to enhance their efficiency and accuracy. Existing models are mainly focusing on supervised learning, but it is time-consuming and challenging to label all samples in AQA dataset. Therefore, semi-supervised learning and unsupervised learning can be conducted more in future research. In addition, due to the rapid development of transformer in both image and video domains (Liu et al., 2022; Vaswani et al., 2017), future research can delve further into its ability to capture motion features in AQA tasks. Furthermore, most existing studies assess the action quality by using C3D, P3D and I3D to extract features from RGB videos, which makes features contain ambiguous scene information and ignore the internal connections between joints. In contrast, GCN can extract features from skeleton data which contains intuitive structural body pose and joint information. However, acquiring accurate posture from motion cameras and pose estimators is challenging in fast movements and occluded situations. To overcome this challenge, although some studies (Gao, Pan et al., 2023; Nekoui et al., 2021; Pan et al., 2019) combine

Table 13
Overview of all identified papers. (CA = Classification accuracy, PCC = Pearson correlation coefficient, SRC = Spearman correlation coefficient, MED = Mean euclidean distance, MRC = Mean rank correlation, KC = Kendall correlation, MSE = Mean square error, MAE = Mean absolute error, ED = Euclidean distance, R-ℓ_2 = Relative ℓ_2-distance), ICC = The Intra-group Correlation Coefficient.

#	Ref.	Application	Dataset	Data	Methods	Evaluation metrics
1	Çeliktutan et al. (2013)	Physical exercise	Perturbed Workout SU-10 Gesture	Skeleton	SVM	CA
2	Venkataraman et al. (2013)	Stroke rehabilitation	Stroke Rehabilitation	Skeleton	SVM	PCC
3	Venkataraman et al. (2014)	Stroke rehabilitation	self-created	Skeleton	Decision Tree	PCC
4	Pirsiavash et al. (2014)	Olympic games	MIT-Diving MIT-Skating	Skeleton	Pose+DCT+SVR	MRC
5	Jin et al. (2016)	Fitness	self-created	Skeleton	DTW	ED
6	Venkataraman et al. (2016)	Stroke rehabilitation	self-created	Skeleton	linear model	PCC
7	Venkataraman and Turaga (2016)	Stroke rehabilitation	Stroke Rehabilitation	Skeleton	SVM	PCC
8	Parmar and Tran Morris (2017)	Olympic games	MIT-Diving MIT-Skating UNLV-Vault UNLV-Divng	Video	C3D-SVR C3D-LSTM	SRC
9	Li et al. (2018)	Olympic games	UNLV-Diving UNLV-Vault MIT-Skating	Video	C3D	SRC, MED
10	Xiang et al. (2018)	Olympic games	UNLV-Dive	Video	ED-TCN+P3D+FC/LR/SVR	SRC
11	Parmar and Morris (2019b)	Olympic games	MTL-AQA	Video	C3D	SRC
12	Li, Chai et al. (2019)	Olympic games	MIT-Diving UNLV-Diving UNLV-Vault	Video	C3D	SRC, MED
13	Pan et al. (2019)	Olympic games surgical skill training	AQA-7 Dataset JIGSAWS Dataset	VideoSkeleton	I3D	SRC
14	Parmar and Morris (2019a)	Olympic games	AQA-7 Dataset	Video	C3D-LSTM	SRC
15	Dajime et al. (2020)	Physical exercise	self-created	Skeleton	MLR	CA
16	Yu et al. (2020)	Rehabilitation exercise	UI-PRMD	Skeleton	GCN	CA
17	Wang, Jin et al. (2020)	Hand skill	Origami Video	Skeleton	LSTM, DCT+SVC, DFT+SVC	CA
18	Tang et al. (2020)	Olympic games surgical skill training	AQA-7 MTL-AQA JIGSAWS	Video	I3D	SRC
19	Zeng et al. (2020)	Olympic games	Rhythmic Gymnastics MIT-Skating	Video	I3D, ResNet, GCN, Attention	SRC
20	Lei et al. (2020)	Olympic games	MIT-Diving MIT-Skating UNLV vault	Skeleton	support vector regression ridge regression	MRC
21	Gao et al. (2020)	Surgical skill training	JIGSAWS TASD-2	Video	I3D	SRC
22	Wang, Du et al. (2020)	Olympic games	UNLV-Skating UNLV-Vault UNLV-Divng	Video	SCN, TCN, Attention	SRC
23	Nekoui et al. (2021)	Olympic games	AQA-7	VideoSkeleton	I3D, HRNet, Attention, TCN, SCN	SRC
24	Yu, Rao et al. (2021)	Olympic games surgical skill training	AQA-7 MTL-AQA JIGSAWS	Video	I3D	SRC
25	Nagai et al. (2021)	Olympic games	MTL-AQA	Video	C3D	SRC
26	Parmar et al. (2021)	Piano skill	PISA	VideoAudio	3DCNN, ResNet-18	CA
27	Li, Lei et al. (2021)	Olympic games	MIT-Skating	Skeleton	ST-GCN	SRC, MED
28	Dong et al. (2021)	Olympic games	UNLV-Diving	Video	ED-TCN, P3D	SRC
29	Wang, Yang, Zhai, Chen et al. (2021)	Olympic games	AQA-7 MTL-AQA	Video	Self-Attention, I3D	SRC
30	Roditakis et al. (2021)	Olympic games	MTL-AQA	Video	I3D, Temporal Cycle-Consistency (TCC)	SRC
31	Lei et al. (2021)	Olympic games	AQA-7	Video	I3D, Attention	SRC, MED
32	Li, Bhat et al. (2021)	Behavioural therapies	Play Therapy 13	Video	I3D	SRC, MSE
33	Yu, Liu et al. (2021)	Behavioural therapies	UI-PRMD EHE dataset	Skeleton	2T-GCN	SRC, ED
34	Li, Hu et al. (2021)	Fitness actions	Fitness-28 dataset	SkeletonDepth	Iterative Closest Point (ICP)	SRC
35	Farabi et al. (2022)	Olympic games	MTL-AQA	Video	3D and (2+1)D ResNets	SRC
36	Freire-Obregón et al. (2022)	Ultra-running	TGC	Video	I3D	CA
37	Xu, Rao et al. (2022)	Olympic games	FineDiving	Video	I3D	SRC, R-ℓ_2
38	Zhang, Dong et al. (2022)	Sports event	UNLV-Diving	Video	P3D	SRC, MED, MSE
39	Li, Zhang, Dong et al. (2022)	Olympic games	AQA-7 MTL-AQA	Video	ResNet, Temporal Encoder	SRC
40	Li, Zhang, Lei et al. (2022)	Olympic games	AQA-7 MTL-AQA	Video	ResNet, Temporal encoder	SRC
41	Bai et al. (2022)	Olympic games	MTL-AQA AQA-7 JIGSAW	Video	I3D, Transformer	SRC, R-ℓ_2

(continued on next page)

Table 13 (*continued*).

42	Wang et al. (2022)	Tai Chi Quan gesture	UMONS-TAICHI Walking Gait	Skeleton	LSTM	SRC, CA, PCC
43	Zhang, Pan et al. (2022)	Olympic games	MTL-AQA Rhythmic Gymnastics	Video	I3D	SRC
44	Baby et al. (2022)	Surgical skill training	JIGSAWS Nets	Video	TCN	SRC
45	Pan et al. (2022)	Olympic games surgical skill training	AQA-7 JIGSAWS	Video	I3D	SRC
46	Li, Lei et al. (2022)	Olympic games	MIT-Skating FIS-V	Skeleton	ST-GCN, LSTM	SRC, MSE
47	Zhang, Xiong et al. (2022)	Olympic games	MTL-AQA	Video	I3D	SRC
48	Li, Cui et al. (2022)	Olympic games	FineGym dataset	Video	C3D, MLP, VAE, Transformer	MSE
49	Xu, Zeng et al. (2022)	Olympic games	Rhythmic Gymnastics Fis-V	Video	VST	SRC
50	Li, Hu et al. (2022)	Tai Chi Quan gesture	TaiChi-24	Skeleton	ST-GCN	CA
51	Zhang, Wang et al. (2023)	Olympic games	AQA-7	Skeleton	GCN, TCN	SRC
52	Gao, Pan et al. (2023)	Olympic games surgical skill training daily actions	JIGSAWS TASD-2 PaSk AQA-7 EPIC-Skills BEST	VideoSkeleton	I3D, Attention	SRC
53	Liu, Wang et al. (2023)	Pull-ups test	self-created	Skeleton	PEPoseNet, Random Forest	CA
54	Freire-Obregon et al. (2023)	Ultra-running	TGC20ReID	video	X3D	MAE
55	Joung et al. (2023)	Fitness	SQUAT AI-HUB FITNESS	skeleton	GCN	F1
56	Lei et al. (2023)	Olympic games	MIT-Skating Rhythmic Gymnastics	Skeleton	GCN	SRC
57	Ji et al. (2023)	Olympic games	FineFS (proposed) Fis-V	Video	VST	SRC
58	Liu, Cheng et al. (2023)	Olympic games	RFSJ	Video	I3D, Transformer, Attention	SRC, R-ℓ_2
59	Kanade et al. (2023a)	Rehabilitation exercise	UI-PRMD KIMORE	Skeleton	CNN, Attention	MAE
60	Sun, Hu et al. (2023)	Olympic games Surgical skill training	AQA-7 JIGSAWS	Video	RNN, Attention	MAE
61	Kanade et al. (2023b)	Rehabilitation exercise	KIMORE	Skeleton	CNN, LSTM	RMSE
62	Li, Ling et al. (2023)	Rehabilitation exercise	UI-PRMD IRDS	Skeleton	GCN	CA
63	Chariar et al. (2023)	Physical exercise	self-created	Skeleton	LSTM, Attention	CA
64	Fang, Zhou et al. (2023)	Olympic games	MTL-AQA FineDiving	Video	I3D, Transformer	SRC, R-ℓ_2
65	Hipiny et al. (2023)	Dance	self-created	Skeleton	pairwise	Precision
66	Li, Zhang et al. (2023)	Olympic games	AQA-7 MTL-AQA	Video	ResNet, Temporal Encoder	SRC
67	Lian and Shao (2023)	Olympic games	FineDiving	Video	(2+1)D ResNet	SRC, R-ℓ_2
68	Mourchid and Slama (2023)	Rehabilitation exercise	UI-PRMD	Skeleton	GCN, Attention	MAE
69	Fang, Luo et al. (2023)	Rehabilitation exercise	self-created	Skeleton	Threshold	ICC
70	Matsuyama et al. (2023)	Olympic games	self-created	Video	CNN, TCN	SRC, PCC
71	Zhou, Cai et al. (2023)	Juvenile dermatomyositis	JDM dataset	Video	I3D	SRC, R-ℓ_2
72	Li, Tian et al. (2023)	Martial arts	Fri2023	Skeleton	Transformer	CA
73	Wang et al. (2023)	Fitness	self-created	Skeleton	DTW	–
74	Ingwersen et al. (2023)	Golf skill	self-created	Video	3D ResNet	MSE, KC
75	Du et al. (2023)	Olympic games	FS1000 Fis-v MTL-AQA OlympicFS	Video	VST, Transformer	MSE, SRC
76	Gedamu et al. (2023)	Olympic games	FineDiving AQA-7 MTL-AQA	Video	I3D, Transformer	SRC, R-ℓ_2
77	Hirosawa et al. (2023)	Olympic games	self-created	VideoSkeleton	VGG16	RMSE, SRC, MAE
78	Liu, Zhai et al. (2023)	Olympic games	FineDiving	Video	I3D	SRC
79	Li, Chheang et al. (2023)	Behaviour therapy	MMASD	–	–	–
80	Zhang, Dong et al. (2023)	Olympic games	UNLV-Diving	Video	P3D	SRC, MSE, MED
81	Zhang, Zhou et al. (2023)	Behaviour therapy	DREAM	Skeleton	CNN, LSTM	SRC
82	Zhou, Ma et al. (2023)	Olympic games surgical skill training	MTL-AQA JIGSAWS AQA-7	Video	I3D, GCN	SRC, R-ℓ_2
83	Zhang, Dai et al. (2023)	Olympic games	LOGO	Video	VST, GCN	SRC, R-ℓ_2
84	Wang et al. (2024)	Medical skill training	CPR-Coach	Video	ResNet, Transformer	mAP, mmit mAP
85	Yuan (2024)	Wushu teaching	Taiji	Video	CNN	MSE
86	Huang and Li (2024)	Olympic games	UNLV-Dive AQA-7	Video	I3D, TCN	SRC

(*continued on next page*)

Table 13 (*continued*).

87	Dadashzadeh et al. (2024)	Olympic games surgical skill training rehabilitation	JIGSAWS MTL-AQA FineDiving PD4T	Video	I3D	SRC
88	Gan et al. (2024)	Olympic games	SkatingVerse	Video	–	SRC
89	Zhang, Pan et al. (2024)	Skill training	EPIC-Skill JIGSAWS BEST AQA-7	Video	I3D, Attention	CA, SRC
90	Ke et al. (2024)	Olympic games surgical skill training	JIGSAWS MTL-AQA FineDiving AQA-7	Video	I3D, Transformer	SRC, $R\text{-}\ell_2$
91	Yu et al. (2024)	Rehabilitation exercise	UI-PRMD KIMORE EHE	Skeleton	GCN	SRC, ED
92	Sardari et al. (2024)	Rehabilitation exercise	UI-PRMD KIMORE	Skeleton	TCN	MAE, RMSE
93	Nagai et al. (2024)	Windsurfing	MMW-AQA	VideoIMU GPS	Transformer	SRC
94	He et al. (2024)	Olympic games	FineDiving	Video	ResNet, X3D	SRC, $R\text{-}\ell_2$
95	Zhang, Chen et al. (2024)	Olympic games surgical skill training	AQA-7 MTL-AQA JIGSAWS	Video	I3D	SRC
96	Zeng and Zheng (2024)	Olympic games	RG, Fis-V	Video	VST, I3D, AST, UNMT, MAST	SRC

skeleton information into RGB-based methods in AQA, they do not explore the fusion of multi-modal in features level. Recently, some studies (Bruce, Liu, Zhang, Zhong, & Chan, 2022; Das, Dai, Yang, & Bremond, 2021) propose multi-modal methods by integrating skeleton and RGB data in the feature level for human action recognition and achieve remarkable performance. Future direction can be devoted to exploring the fusion of RGB and other useful information such as skeleton, saliency, audio etc. in AQA tasks. There are both similarities and distinctions in the evaluation of different actions and skills. Currently, AQA models are trained individually for each specific action, resulting in separate models for each type of evaluation. This approach tends to ignore the potential similarities that exist across different actions and skills, such as common movement patterns, evaluation criteria, or similar performance attributes. Recognizing and utilizing these similarities could lead to more efficient training processes and enhance the models' ability to generalize across various types of actions and skills. With the development of large models in the computer vision community (Kirillov et al., 2023), future research could focus on developing methods to systematically identify and incorporate these similarities, thereby improving the generalization capabilities of AQA models. This might involve exploring cross-domain features, designing unified frameworks for diverse evaluations, or employing advanced algorithms to transfer knowledge between related tasks. Enhancing the generalization capabilities of AQA models in this way could lead to more adaptable and effective systems that perform well across a broader range of assessments of actions and skills.

6. Conclusion

In this paper, we systematically reviewed 96 research articles to provide a comprehensive and critical overview of vision-based human AQA. We included 96 research articles published in a journal or conference in Scopus, IEEE Xplore and Web of Science until July 2024 to present a comprehensive review of application, data acquisition, datasets, skeleton-based methods, RGB video-based methods and evaluation metrics in AQA. We have observed due to the limitation of datasets, most existing AQA studies focused on Olympic event scoring, while a small number of studies focused on healthcare and other domains. In addition, we summarized existing AQA methods into skeletal-based and video-based. Furthermore, we found that the evaluation metrics for AQA methods vary across different studies due to diverse definitions of AQA tasks. The most commonly used evaluation metric among these studies is SRC and we have provided the SRC performance of some state-of-the-art methods on three commonly used datasets: MTL-AQA, AQA-7 and JIGSAW datasets. Finally, we analysed the current challenges and provided suggestions for future research directions. This systematic review can be a helpful guide for researchers to explore recent literature, public datasets, and state-of-the-art methods in AQA.

CRediT authorship contribution statement

Jiang Liu: Methodology, Writing – original draft. **Huasheng Wang:** Methodology, Writing – review & editing. **Katarzyna Stawarz:** Conceptualization, Writing – review & editing. **Shiyin Li:** Supervision. **Yao Fu:** Methodology. **Hantao Liu:** Conceptualization, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work is funded by the China Scholarship Council - ID 20220642 0009.

Appendix

See Table 13.

Data availability

Data will be made available on request.

References

Ahad, M. A. R., Antar, A. D., & Shahid, O. (2019). Vision-based action understanding for assistive healthcare: A short review. In *CVPR workshops* (pp. 1–11).

Al-Faris, M., Chiverton, J., Ndzi, D., & Ahmed, A. (2020). A review on computer vision-based methods for human action recognition. *Journal of Imaging*, *6*(6).

Baby, B., Chasmai, M., Banerjee, T., Suri, A., Banerjee, S., & Arora, C. (2022). Representation learning using rank loss for robust neurosurgical skills evaluation. In *Proceedings - international conference on image processing* (pp. 4048–4052).

Baca, A., Dabnichki, P., Hu, C.-W., Kornfeind, P., & Exel, J. (2022). Ubiquitous computing in sports and physical activity—Recent trends and developments. *Sensors*, *22*(21), 8370.

Bai, Y., Zhou, D., Zhang, S., Wang, J., Ding, E., Guan, Y., et al. (2022). Action quality assessment with temporal parsing transformer. In *LNCS*: *Vol. 13664*, *Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics)* (pp. 422–438).

Bazarevsky, V., Grishchenko, I., Raveendran, K., Zhu, T., Zhang, F., & Grundmann, M. (2020). Blazepose: On-device real-time body pose tracking. arXiv preprint arXiv:2006.10204.

Besl, P. J., & McKay, N. D. (1992). Method for registration of 3-D shapes. *Vol. 1611*, In *Sensor fusion IV: control paradigms and data structures* (pp. 586–606). SPIE.

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and regression trees–CRC press*. Boca Raton, Florida.

Bruce, X., Liu, Y., Zhang, X., Zhong, S.-h., & Chan, K. C. (2022). Mmnet: A model-based multimodal network for human action recognition in rgb-d videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(3), 3522–3538.

Brunnström, K., Djupsjöbacka, A., Ozolins, O., Billingham, J., Wistel, K., & Evans, N. (2023). Quality measurement methods for video assisting refereeing systems. *Sports Engineering*, *26*(1), 17.

Cao, Z., Simon, T., Wei, S.-E., & Sheikh, Y. (2017). Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7291–7299).

Carreira, J., & Zisserman, A. (2017). Quo vadis, action recognition? A new model and the kinetics dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 6299–6308).

Çeliktutan, O., Akgül, C., Wolf, C., & Sankur, B. (2013). Graph-based analysis of physical exercise actions. In *MIIRH 2013 - proceedings of the 1st ACM international workshop on multimedia indexing and information retrieval for heathcare, co-located with ACM multimedia 2013* (pp. 23–31).

Chariar, M., Rao, S., Irani, A., Suresh, S., & Asha, C. (2023). AI trainer: Autoencoder based approach for squat analysis and correction. *IEEE Access*, *11*, 107135–107149.

Chen, W., Yu, C., Tu, C., Lyu, Z., Tang, J., Ou, S., et al. (2020). A survey on hand pose estimation with wearable sensors and computer-vision-based methods. *Sensors*, *20*(4), 1074.

Dadashzadeh, A., Duan, S., Whone, A., & Mirmehdi, M. (2024). Pecop: Parameter efficient continual pretraining for action quality assessment. In *Proceedings - 2024 IEEE winter conference on applications of computer vision* (pp. 42–52).

Dajime, P., Smith, H., & Zhang, Y. (2020). Automated classification of movement quality using the microsoft kinect V2 sensor. *Computers in Biology and Medicine*, *125*.

Das, S., Dai, R., Yang, D., & Bremond, F. (2021). Vpn++: Rethinking video-pose embeddings for understanding activities of daily living. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(12), 9703–9717.

Do, H. M., Welch, K. C., & Sheng, W. (2021). Soham: A sound-based human activity monitoring framework for home service robots. *IEEE Transactions on Automation Science and Engineering*, *19*(3), 2369–2383.

Dong, L.-J., Zhang, H.-B., Shi, Q., Lei, Q., Du, J.-X., & Gao, S. (2021). Learning and fusing multiple hidden substages for action quality assessment. *Knowledge-Based Systems*, *229*.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

Du, Z., He, D., Wang, X., & Wang, Q. (2023). Learning semantics-guided representations for scoring figure skating. *IEEE Transactions on Multimedia*, *26*, 4987–4997.

Dwibedi, D., Aytar, Y., Tompson, J., Sermanet, P., & Zisserman, A. (2019). Temporal cycle-consistency learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 1801–1810).

Faller, A. J. (1981). An average correlation coefficient. *Journal of Applied Meteorology (1962-1982)*, 203–205.

Fang, Y., Luo, Z., Huang, F., Wang, Z., Li, D., & Hua, X. (2023). Developing a mixed reality-based game for post-stroke motor rehabilitation: Combining training and assessment. In *2023 9th international conference on virtual reality* (pp. 393–399).

Fang, H., Zhou, W., & Li, H. (2023). End-to-end action quality assessment with action parsing transformer. In *2023 IEEE international conference on visual communications and image processing*.

Farabi, S., Himel, H., Gazzali, F., Hasan, M., Kabir, M., & Farazi, M. (2022). Improving action quality assessment using weighted aggregation. In *LNCS*: *Vol. 13256*, *Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics)* (pp. 576–587).

Freire-Obregón, D., Lorenzo-Navarro, J., & Castrillón-Santana, M. (2022). Decontextualized I3D ConvNet for ultra-distance runners performance analysis at a glance. In *LNCS*: *Vol. 13233*, *Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics)* (pp. 242–253).

Freire-Obregon, D., Lorenzo-Navarro, J., Santana, O. J., Hernandez-Sosa, D., & Castrillon-Santana, M. (2023). An X3D neural network analysis for runner's performance assessment in a wild sporting environment. In *Proceedings of MVA 2023 - 18th international conference on machine vision and applications*.

Gan, Z., Jin, L., Cheng, Y., Cheng, Y., Teng, Y., Li, Z., et al. (2024). SkatingVerse: A large-scale benchmark for comprehensive evaluation on human action understanding. *IET Computer Vision*.

Gao, Z., Cui, X., Zhuo, T., Cheng, Z., Liu, A.-A., Wang, M., et al. (2023). A multitemporal scale and spatial–temporal transformer network for temporal action localization. *IEEE Transactions on Human-Machine Systems*, *53*(3), 569–580.

Gao, J., Pan, J.-H., Zhang, S.-J., & Zheng, W.-S. (2023). Automatic modelling for interactive action assessment. *International Journal of Computer Vision*, *131*(3), 659–679.

Gao, Y., Vedula, S. S., Reiley, C. E., Ahmidi, N., Varadarajan, B., Lin, H. C., et al. (2014). Jhu-isi gesture and skill assessment working set (jigsaws): A surgical activity dataset for human motion modeling. *Vol. 3*, In *MICCAI workshop: M2cai*.

Gao, J., Zheng, W.-S., Pan, J.-H., Gao, C., Wang, Y., Zeng, W., et al. (2020). An asymmetric modeling for action assessment. In *LNCS*: *Vol. 12375*, *Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics)* (pp. 222–238).

Gedamu, K., Ji, Y., Yang, Y., Shao, J., & Shen, H. T. (2023). Fine-grained spatio-temporal parsing network for action quality assessment. *IEEE Transactions on Image Processing*, *32*, 6386–6400.

Gharasuie, M. M., Jennings, N., & Jain, S. (2021). Performance monitoring for exercise movements using mobile cameras. In *Proceedings of the workshop on body-centric computing systems* (pp. 1–6).

Gkioxari, G., & Malik, J. (2015). Finding action tubes. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 759–768).

Guo, Y., Yang, C., Rao, A., Liang, Z., Wang, Y., Qiao, Y., et al. (2024). AnimateDiff: Animate your personalized text-to-image diffusion models without specific tuning. In *International conference on learning representations*.

Hao, Y., Shi, Z., & Liu, Y. (2022). WiFi-based spatiotemporal human action perception. In *2022 IEEE international conference on image processing* (pp. 3581–3585). IEEE.

He, T., Liu, H., Li, Y., Ma, X., Zhong, C., Zhang, Y., et al. (2024). Collaborative weakly supervised video correlation learning for procedure-aware instructional video analysis. *Vol. 38*, In *Proceedings of the AAAI conference on artificial intelligence* (pp. 2112–2120).

Hipiny, I., Ujir, H., Alias, A., Shanat, M., & Ishak, M. (2023). Who danced better? Ranked tiktok dance video dataset and pairwise action quality assessment method. *International Journal of Advances in Intelligent Informatics*, *9*(1), 96–107.

Hirosawa, S., Kato, T., Yamashita, T., & Aoki, Y. (2023). Action quality assessment model using specialists' gaze location and kinematics data—Focusing on evaluating figure skating jumps. *Sensors*, *23*(22).

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780.

Hong, W., Ding, M., Zheng, W., Liu, X., & Tang, J. (2022). CogVideo: Large-scale pretraining for text-to-video generation via transformers. arXiv preprint arXiv:2205.15868.

Hou, R., Chen, C., & Shah, M. (2017). Tube convolutional neural network (T-CNN) for action detection in videos. In *Proceedings of the IEEE international conference on computer vision* (pp. 5822–5831).

Huang, F., & Li, J. (2024). Assessing action quality with semantic-sequence performance regression and densely distributed sample weighting. *Applied Intelligence: The International Journal of Artificial Intelligence, Neural Networks, and Complex Problem-Solving Technologies*, *54*(4), 3245–3259.

Huang, W., Yang, J., Luo, H., & Zhang, H. (2023). Human table tennis actions recognition and evaluation method based on skeleton extraction. In *2023 3rd international conference on consumer electronics and computer engineering* (pp. 7–13).

Ingwersen, C. K., Xarles, A., Clapes, A., Madadi, M., Jensen, J. N., Hannemose, M. R., et al. (2023). Video-based skill assessment for golf: Estimating golf handicap. In *MMSports 2023 - proceedings of the 6th international workshop on multimedia content analysis in sports, co-located with: MM 2023* (pp. 31–39).

Islam, M. M., Nooruddin, S., Karray, F., & Muhammad, G. (2022). Human activity recognition using tools of convolutional neural networks: A state of the art review, data sets, challenges, and future prospects. *Computers in Biology and Medicine*, *149*, Article 106060.

Jain, A., Tompson, J., LeCun, Y., & Bregler, C. (2015). Modeep: A deep learning framework using motion features for human pose estimation. In *Computer vision–ACCV 2014: 12th Asian conference on computer vision, Singapore, Singapore, November 1-5, 2014, revised selected papers, part II 12* (pp. 302–315). Springer.

Jakab, S., Davis, P., & Whyte, I. (2023). An exploratory investigation of traditional scoring in diving and relationships to the development of artificial intelligence opportunities. *Scientific Journal of Sport and Performance*, *2*(3), 300–313.

Ji, Y., Ye, L., Huang, H., Mao, L., Zhou, Y., & Gao, L. (2023). Localization-assisted uncertainty score disentanglement network for action quality assessment. In *MM 2023 - proceedings of the 31st ACM international conference on multimedia* (pp. 8590–8597).

Jin, X., Yao, Y., Jiang, Q., Huang, X., Zhang, J., Zhang, X., et al. (2016). Virtual personal trainer via the kinect sensor. *Vol. 2016-February*, In *International conference on communication technology proceedings* (pp. 460–463).

Jo, B., & Kim, S. (2022). Comparative analysis of OpenPose, PoseNet, and MoveNet models for pose estimation in mobile devices. *Traitement du Signal*, *39*(1), 119–124.

Joung, C.-I., Byun, S., & Baek, S. (2023). Contrastive learning for action assessment using graph convolutional networks with augmented virtual joints. *IEEE Access*, *11*, 88895–88907.

Kanade, A., Sharma, M., & Muniyandi, M. (2023a). Attention-guided deep learning framework for movement quality assessment. *Vol. 2023-June*, In *ICASSP, IEEE international conference on acoustics, speech and signal processing - proceedings*.

Kanade, A., Sharma, M., & Muniyandi, M. (2023b). Tele-EvalNet: A low-cost, teleconsultation system for home based rehabilitation of stroke survivors using multiscale CNN-convlstm architecture. In *LNCS*: *Vol. 13806*, *Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics)* (pp. 738–750).

Karayaneva, Y., Sharifzadeh, S., Jing, Y., Chetty, K., & Tan, B. (2019). Sparse feature extraction for activity detection using low-resolution IR streams. In *2019 18th IEEE international conference on machine learning and applications* (pp. 1837–1843). IEEE.

Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., & Fei-Fei, L. (2014). Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1725–1732).

Ke, X., Xu, H., Lin, X., & Guo, W. (2024). Two-path target-aware contrastive regression for action quality assessment. *Information Sciences*, *664*.

Kim, J.-K., Lee, K., & Hong, S. G. (2023). Detection of important features and comparison of datasets for fall detection based on wrist-wearable devices. *Expert Systems with Applications*, *234*, Article 121034.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., et al. (2023). Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 4015–4026).

Kitchenham, B., Charters, S., et al. (2007). Guidelines for performing systematic literature reviews in software engineering.

Kong, Y., & Fu, Y. (2022). Human action recognition and prediction: A survey. *International Journal of Computer Vision*, *130*(5), 1366–1401.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, *25*.

Lan, G., Wu, Y., Hu, F., & Hao, Q. (2023). Vision-based human pose estimation via deep learning: A survey. *IEEE Transactions on Human-Machine Systems*, *53*(1), 253–268.

Lea, C., Flynn, M. D., Vidal, R., Reiter, A., & Hager, G. D. (2017). Temporal convolutional networks for action segmentation and detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 156–165).

Lea, C., Vidal, R., Reiter, A., & Hager, G. D. (2016). Temporal convolutional networks: A unified approach to action segmentation. In *Computer vision–ECCV 2016 workshops: Amsterdam, the Netherlands, October 8-10 and 15-16, 2016, proceedings, part III 14* (pp. 47–54). Springer.

LeCun, Y., Bengio, Y., et al. (1995). Convolutional networks for images, speech, and time series. *The Handbook of Brain Theory and Neural Networks*, *3361*(10), 1995.

Lei, Q., Du, J.-X., Zhang, H.-B., Ye, S., & Chen, D.-S. (2019). A survey of vision-based human action evaluation methods. *Sensors*, *19*(19), 4129.

Lei, Q., Li, H., Zhang, H., Du, J., & Gao, S. (2023). Multi-skeleton structures graph convolutional network for action quality assessment in long videos. *Applied Intelligence: The International Journal of Artificial Intelligence, Neural Networks, and Complex Problem-Solving Technologies*, *53*(19), 21692–21705.

Lei, Q., Zhang, H., & Du, J. (2021). Temporal attention learning for action quality assessment in sports video. *Signal, Image and Video Processing*, *15*(7), 1575–1583.

Lei, Q., Zhang, H.-B., Du, J.-X., Hsiao, T.-C., & Chen, C.-C. (2020). Learning effective skeletal representations on RGB video for fine-grained human action quality assessment. *Electronics (Switzerland)*, *9*(4).

Li, J., Bhat, A., & Barmaki, R. (2021). Improving the movement synchrony estimation with action quality assessment in children play therapy. In *ICMI 2021 - proceedings of the 2021 international conference on multimodal interaction* (pp. 397–406).

Li, Y., Chai, X., & Chen, X. (2018). End-to-end learning for action quality assessment. In *LNCS*: *Vol. 11165*, *Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics)* (pp. 125–134).

Li, Y., Chai, X., & Chen, X. (2019). ScoringNet: Learning key fragment for action quality assessment with ranking loss in skilled sports. In *LNCS*: *Vol. 11366*, *Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics)* (pp. 149–164).

Li, J., Chheang, V., Kullu, P., Brignac, E., Guo, Z., Bhat, A., et al. (2023). MMASD: A multimodal dataset for autism intervention analysis. In *ACM international conference proceeding series* (pp. 397–405).

Li, J., Cui, H., Guo, T., Hu, Q., & Shen, Y. (2020). Efficient fitness action analysis based on spatio-temporal feature encoding. In *2020 IEEE international conference on multimedia & expo workshops* (pp. 1–6). IEEE.

Li, Q., Cui, Z., Kitahara, I., & Sagawa, R. (2022). Precise gymnastic scoring from TV playback. In *GCCE 2022 - 2022 IEEE 11th global conference on consumer electronics* (pp. 412–415).

Li, X., He, Y., & Jing, X. (2019). A survey of deep learning-based human activity recognition in radar. *Remote Sensing*, *11*(9), 1068.

Li, J., Hu, Q., Guo, T., Wang, S., & Shen, Y. (2021). What and how well you exercised? An efficient analysis framework for fitness actions. *Journal of Visual Communication and Image Representation*, *80*.

Li, J., Hu, H., Xing, Q., Wang, X., Li, J., & Shen, Y. (2022). Tai chi action quality assessment and visual analysis with a consumer RGB-D camera. In *2022 IEEE 24th international workshop on multimedia signal processing*.

Li, H.-Y., Lei, Q., Zhang, H.-B., & Du, J.-X. (2021). Skeleton based action quality assessment of figure skating videos. In *Proceedings - 11th international conference on information technology in medicine and education* (pp. 196–200).

Li, H., Lei, Q., Zhang, H., Du, J., & Gao, S. (2022). Skeleton-based deep pose feature learning for action quality assessment on figure skating videos. *Journal of Visual Communication and Image Representation*, *89*.

Li, D., Li, J., Li, H., Niebles, J. C., & Hoi, S. C. (2022). Align and prompt: Video-and-language pre-training with entity prompts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 4953–4963).

Li, C., Ling, X., & Xia, S. (2023). A graph convolutional siamese network for the assessment and recognition of physical rehabilitation exercises. In *LNCS*: *Vol. 14257*, *Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics)* (pp. 229–240).

Li, M., Tian, F., & Li, Y. (2023). Martial arts scoring system based on U-shaped networkwushu intelligent scoring systemlearning to score Chinese wushu. In *ACM international conference proceeding series* (pp. 1441–1446).

Li, M.-Z., Zhang, H.-B., Dong, L.-J., Lei, Q., & Du, J.-X. (2022). Gaussian guided frame sequence encoder network for action quality assessment. *Complex and Intelligent Systems*.

Li, M.-Z., Zhang, H.-B., Dong, L.-J., Lei, Q., & Du, J.-X. (2023). Gaussian guided frame sequence encoder network for action quality assessment. *Complex and Intelligent Systems*, *9*(2), 1963–1974.

Li, M., Zhang, H.-B., Lei, Q., Fan, Z., Liu, J., & Du, J.-X. (2022). Pairwise contrastive learning network for action quality assessment. In *LNCS*: *Vol. 13664*, *Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics)* (pp. 457–473).

Lian, P.-X., & Shao, Z.-G. (2023). Improving action quality assessment with across-staged temporal reasoning on imbalanced data. *Applied Intelligence: The International Journal of Artificial Intelligence, Neural Networks, and Complex Problem-Solving Technologies*, *53*(24), 30443–30454.

Liang, J., Luo, J., Gao, W., & Lu, L. (2021). Research on fitness action evaluation system based on skeleton. In *2021 international conference on computer network, electronic and automation* (pp. 69–74).

Liao, C.-C., Hwang, D.-H., Wu, E., & Koike, H. (2023). AI coach: A motor skill training system using motion discrepancy detection. In *Proceedings of the augmented humans international conference 2023* (pp. 179–189). New York, NY, USA: Association for Computing Machinery.

Liao, Y., Vakanski, A., Xian, M., Paul, D., & Baker, R. (2020). A review of computational approaches for evaluation of rehabilitation exercises. *Computers in Biology and Medicine*, *119*, Article 103687.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*.

Liu, Y., Cheng, X., & Ikenaga, T. (2023). A figure skating jumping dataset for replay-guided action quality assessment. In *MM 2023 - proceedings of the 31st ACM international conference on multimedia* (pp. 2437–2445).

Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S., et al. (2022). Video swin transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 3202–3211).

Liu, H., Simonyan, K., & Yang, Y. (2018). Darts: Differentiable architecture search. arXiv preprint arXiv:1806.09055.

Liu, G., Wang, J., Zhang, Z., Liu, Q., Ren, Y., Zhang, M., et al. (2023). A novel model for intelligent pull-ups test based on key point estimation of human body and equipment. *Mobile Information Systems*, *2023*.

Liu, L., Zhai, P., Zheng, D., & Fang, Y. (2023). Multi-stage action quality assessment method. In *ACM international conference proceeding series* (pp. 116–122).

Machlin, S. R., Chevan, J., Yu, W. W., & Zodet, M. W. (2011). Determinants of utilization and expenditures for episodes of ambulatory physical therapy among adults. *Physical Therapy*, *91*(7), 1018–1029.

MacMahon, C., Mascarenhas, D., Plessner, H., Pizzera, A., Oudejans, R., & Raab, M. (2014). *Sports officials and officiating: Science and practice*. Routledge.

Majumder, S., & Kehtarnavaz, N. (2021). Vision and inertial sensing fusion for human action recognition: A review. *IEEE Sensors Journal*, *21*(3), 2454–2467.

Matsuyama, H., Kawaguchi, N., & Lim, B. (2023). IRIS: Interpretable rubric-informed segmentation for action quality assessment. In *International conference on intelligent user interfaces, proceedings IUI* (pp. 368–378).

Mourchid, Y., & Slama, R. (2023). MR-STGN: Multi-residual spatio temporal graph network using attention fusion for patient action assessment. In *2023 IEEE 25th international workshop on multimedia signal processing*.

Muhamada, A., & Mohammed, A. (2021). Review on recent computer vision methods for human action recognition. *Advances in Distributed Computing and Artificial Intelligence Journal*, *10*(4), 361–379.

Nagai, T., Takeda, S., Matsumura, M., Shimizu, S., & Yamamoto, S. (2021). Action quality assessment with ignoring scene context. *Vol. 2021-September*, In *Proceedings - international conference on image processing* (pp. 1189–1193).

Nagai, T., Takeda, S., Suzuki, S., & Seshimo, H. (2024). MMW-AQA: Multimodal in-the-wild dataset for action quality assessment. *IEEE Access*, 1.

Nekoui, M., Cruz, F., & Cheng, L. (2021). EAGLE-eye: Extreme-pose action grader using detail bird's-eye view. In *Proceedings - 2021 IEEE winter conference on applications of computer vision* (pp. 394–402).

Nguyen, T. N., Huynh, H. H., & Meunier, J. (2018). 3D reconstruction with time-of-flight depth camera and multiple mirrors. *IEEE Access*, *6*, 38106–38114.

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., et al. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *International Journal of Surgery*, *88*, Article 105906.

Pan, J.-H., Gao, J., & Zheng, W.-S. (2019). Action assessment by joint relation graphs. *Vol. 2019-October*, In *Proceedings of the IEEE international conference on computer vision* (pp. 6330–6339).

Pan, J.-H., Gao, J., & Zheng, W.-S. (2022). Adaptive action assessment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(12), 8779–8795.

Papandreou, G., Zhu, T., Kanazawa, N., Toshev, A., Tompson, J., Bregler, C., et al. (2017). Towards accurate multi-person pose estimation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4903–4911).

Parmar, P., & Morris, B. (2019a). Action quality assessment across multiple actions. In *2019 IEEE winter conference on applications of computer vision* (pp. 1468–1476). IEEE.

Parmar, P., & Morris, B. (2019b). What and how well you performed? A multitask learning approach to action quality assessment. *Vol. 2019-June*, In *Proceedings of the IEEE computer society conference on computer vision and pattern recognition* (pp. 304–313).

Parmar, P., Reddy, J., & Morris, B. (2021). Piano skills assessment. In *IEEE 23rd international workshop on multimedia signal processing*.

Parmar, P., & Tran Morris, B. (2017). Learning to score olympic events. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops* (pp. 20–28).

Peng, X., & Schmid, C. (2016). Multi-region two-stream R-CNN for action detection. In *Computer vision–ECCV 2016: 14th European conference, Amsterdam, the Netherlands, October 11–14, 2016, proceedings, part IV 14* (pp. 744–759). Springer.

Pirsiavash, H., Vondrick, C., & Torralba, A. (2014). Assessing the quality of actions. In *Computer vision–ECCV 2014: 13th European conference, Zurich, Switzerland, September 6-12, 2014, proceedings, part VI 13* (pp. 556–571). Springer.

Prvu Bettger, J., Liu, C., Gandhi, D. B., Sylaja, P., Jayaram, N., & Pandian, J. D. (2019). Emerging areas of stroke rehabilitation research in low-and middle-income countries: a scoping review. *Stroke*, *50*(11), 3307–3313.

Qiu, Z., Yao, T., & Mei, T. (2017). Learning spatio-temporal representation with pseudo-3d residual networks. In *Proceedings of the IEEE international conference on computer vision* (pp. 5533–5541).

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., et al. (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748–8763). PMLR.

Roditakis, K., Makris, A., & Argyros, A. (2021). Towards improved and interpretable action quality assessment with self-supervised alignment. In *ACM international conference proceeding series* (pp. 507–513).

Sardari, S., Sharifzadeh, S., Daneshkhah, A., Loke, S. W., Palade, V., Duncan, M. J., et al. (2024). LightPRA: A lightweight temporal convolutional network for automatic physical rehabilitation exercise assessment. *Computers in Biology and Medicine*, *173*.

Sardari, S., Sharifzadeh, S., Daneshkhah, A., Nakisa, B., Loke, S. W., Palade, V., et al. (2023). Artificial intelligence for skeleton-based physical rehabilitation action evaluation: A systematic review. *Computers in Biology and Medicine*, Article 106835.

Setiawan, F., Yahya, B. N., Chun, S.-J., & Lee, S.-L. (2022). Sequential inter-hop graph convolution neural network (SIhGCN) for skeleton-based human action recognition. *Expert Systems with Applications*, *195*, Article 116566.

Sigcha, L., Borzì, L., Amato, F., Rechichi, I., Ramos-Romero, C., Cárdenas, A., et al. (2023). Deep learning and wearable sensors for the diagnosis and monitoring of Parkinson's disease: A systematic review. *Expert Systems with Applications*, Article 120541.

Su, C., Wang, C., Gou, S., Chen, J., Tang, W., & Liu, C. (2022). An action recognition method for manual acupuncture techniques using a tactile array finger cot. *Computers in Biology and Medicine*, *148*, Article 105827.

Sun, W., Hu, Y., Zhang, B., Chen, X., Hao, C., & Gao, Y. (2023). A novel blind action quality assessment based on multi-headed GRU network and attention mechanism. *12717*,

Sun, Z., Ke, Q., Rahmani, H., Bennamoun, M., Wang, G., & Liu, J. (2022). Human action recognition from various data modalities: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Sun, Z., Ke, Q., Rahmani, H., Bennamoun, M., Wang, G., & Liu, J. (2023). Human action recognition from various data modalities: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(3), 3200–3225.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2818–2826).

Tang, Y., Ni, Z., Zhou, J., Zhang, D., Lu, J., Wu, Y., et al. (2020). Uncertainty-aware score distribution learning for action quality assessment. In *Proceedings of the IEEE computer society conference on computer vision and pattern recognition* (pp. 9836–9845).

Tits, M., Laraba, S., Caulier, E., Tilmanne, J., & Dutoit, T. (2018). UMONS-TAICHI: A multimodal motion capture dataset of expertise in taijiquan gestures. *Data in Brief*, *19*, 1214–1221.

Toshniwal, D., Patil, A., & Vachhani, N. (2022). AI coach for badminton. In *2022 3rd international conference for emerging technology* (pp. 1–7).

Tran, D., Bourdev, L., Fergus, R., Torresani, L., & Paluri, M. (2015). Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision* (pp. 4489–4497).

Vakanski, A., Jun, H.-p., Paul, D., & Baker, R. (2018). A data set of human body movements for physical rehabilitation exercises. *Data*, *3*(1), 2.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*.

Venkataraman, V., & Turaga, P. (2016). Shape distributions of nonlinear dynamical systems for video-based inference. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *38*(12), 2531–2543.

Venkataraman, V., Turaga, P., Baran, M., Lehrer, N., Du, T., Cheng, L., et al. (2016). Component-level tuning of kinematic features from composite therapist impressions of movement quality. *IEEE Journal of Biomedical and Health Informatics*, *20*(1), 143–152.

Venkataraman, V., Turaga, P., Lehrer, N., Baran, M., Rikakis, T., & Wolf, S. (2013). Attractor-shape for dynamical analysis of human movement: Applications in stroke rehabilitation and action recognition. In *IEEE computer society conference on computer vision and pattern recognition workshops* (pp. 514–520).

Venkataraman, V., Turaga, P., Lehrer, N., Baran, M., Rikakis, T., & Wolf, S. (2014). Decision support for stroke rehabilitation therapy via describable attribute-based decision trees. In *2014 36th annual international conference of the IEEE engineering in medicine and biology society* (pp. 3154–3159).

Wang, J., Du, Z., Li, A., & Wang, Y. (2020). Assessing action quality via attentive spatio-temporal convolutional networks. In *LNCS*: *Vol. 12306*, *Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics)* (pp. 3–16).

Wang, T., Jin, M., Wang, J., Wang, Y., & Li, M. (2020). Towards a data-driven method for RGB video-based hand action quality assessment in real time. In *Proceedings of the ACM symposium on applied computing* (pp. 2117–2120).

Wang, X., Li, J., & Hu, H. (2022). Skeleton-based action quality assessment via partially connected LSTM with triplet losses. In *LNCS*: *Vol. 13536*, *Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics)* (pp. 220–232).

Wang, L., Su, B., Liu, Q., Gao, R., Zhang, J., & Wang, G. (2023). Human action recognition based on skeleton information and multi-feature fusion. *Electronics (Switzerland)*, *12*(17).

Wang, S., Yang, D., Zhai, P., Chen, C., & Zhang, L. (2021). TSA-net: Tube self-attention network for action quality assessment. In *MM 2021 - proceedings of the 29th ACM international conference on multimedia* (pp. 4902–4910).

Wang, S., Yang, D., Zhai, P., Yu, Q., Suo, T., Sun, Z., et al. (2021). A survey of video-based action quality assessment. In *2021 international conference on networking systems of AI* (pp. 1–9). IEEE.

Wang, S., Yang, D., Zhai, P., & Zhang, L. (2024). CPR-CLIP: Multimodal pre-training for composite error recognition in CPR training. *IEEE Signal Processing Letters*, *31*, 211–215.

Xiang, X., Tian, Y., Reiter, A., Hager, G., & Tran, T. (2018). S3D: Stacking segmental P3D for action quality assessment. In *Proceedings - international conference on image processing* (pp. 928–932).

Xu, C., Fu, Y., Zhang, B., Chen, Z., Jiang, Y.-G., & Xue, X. (2019). Learning to score figure skating sport videos. *IEEE Transactions on Circuits and Systems for Video Technology*, *30*(12), 4578–4590.

Xu, J., Rao, Y., Yu, X., Chen, G., Zhou, J., & Lu, J. (2022). FineDiving: A fine-grained dataset for procedure-aware action quality assessment. *Vol. 2022-June*, In *Proceedings of the IEEE computer society conference on computer vision and pattern recognition* (pp. 2939–2948).

Xu, A., Zeng, L.-A., & Zheng, W.-S. (2022). Likert scoring with grade decoupling for long-term action assessment. *Vol. 2022-June*, In *Proceedings of the IEEE computer society conference on computer vision and pattern recognition* (pp. 3222–3231).

Yan, S., Xiong, Y., & Lin, D. (2018). Spatial temporal graph convolutional networks for skeleton-based action recognition. *Vol. 32*, In *Proceedings of the AAAI conference on artificial intelligence*.

Yao, B., & Zhu, S.-C. (2009). Learning deformable action templates from cluttered videos. In *2009 IEEE 12th international conference on computer vision* (pp. 1507–1514). IEEE.

Yin, M., He, S., Soomro, T. A., & Yuan, H. (2023). Efficient skeleton-based action recognition via multi-stream depthwise separable convolutional neural network. *Expert Systems with Applications*, *226*, Article 120080.

Yu, B., Liu, Y., & Chan, K. (2020). Skeleton-based detection of abnormalities in human actions using graph convolutional networks. In *Proceedings - 2020 2nd international conference on transdisciplinary AI* (pp. 131–137).

Yu, B. X., Liu, Y., Chan, K. C., & Chen, C. W. (2024). EGCN++: A new fusion strategy for ensemble learning in skeleton-based rehabilitation exercise assessment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–16.

Yu, B., Liu, Y., Chan, K., Yang, Q., & Wang, X. (2021). Skeleton-based human action evaluation using graph convolutional network for monitoring Alzheimer's progression. *Pattern Recognition*, *119*.

Yu, X., Rao, Y., Zhao, W., Lu, J., & Zhou, J. (2021). Group-aware contrastive regression for action quality assessment. In *Proceedings of the IEEE international conference on computer vision* (pp. 7899–7908).

Yuan, X. (2024). Informatization exploration of wushu teaching management platform in colleges and universities under the concept of modern education. *Applied Mathematics and Nonlinear Sciences*, *9*(1).

Zeng, L.-A., Hong, F.-T., Zheng, W.-S., Yu, Q.-Z., Zeng, W., Wang, Y.-W., et al. (2020). Hybrid dynamic-static context-aware attention network for action assessment in long videos. In *MM 2020 - proceedings of the 28th ACM international conference on multimedia* (pp. 2526–2534).

Zeng, L.-A., & Zheng, W.-S. (2024). Multimodal action quality assessment. *IEEE Transactions on Image Processing*, *33*, 1600–1613.

Zhang, B., Chen, J., Xu, Y., Zhang, H., Yang, X., & Geng, X. (2024). Auto-encoding score distribution regression for action quality assessment. *Neural Computing and Applications*, *36*(2), 929–942.

Zhang, S., Dai, W., Wang, S., Shen, X., Lu, J., Zhou, J., et al. (2023). Logo: A long-form video dataset for group action quality assessment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 2405–2414).

Zhang, H.-B., Dong, L.-J., Lei, Q., Yang, L.-J., & Du, J.-X. (2022). Label-reconstruction-based pseudo-subscore learning for action quality assessment in sporting events. *Applied Intelligence: The International Journal of Artificial Intelligence, Neural Networks, and Complex Problem-Solving Technologies*.

Zhang, H.-B., Dong, L.-J., Lei, Q., Yang, L.-J., & Du, J.-X. (2023). Label-reconstruction-based pseudo-subscore learning for action quality assessment in sporting events. *Applied Intelligence: The International Journal of Artificial Intelligence, Neural Networks, and Complex Problem-Solving Technologies*, *53*(9), 10053–10067.

Zhang, S.-J., Pan, J.-H., Gao, J., & Zheng, W.-S. (2022). Semi-supervised action quality assessment with self-supervised segment feature recovery. *IEEE Transactions on Circuits and Systems for Video Technology*, *32*(9), 6017–6028.

Zhang, S.-J., Pan, J.-H., Gao, J., & Zheng, W.-S. (2024). Adaptive stage-aware assessment skill transfer for skill determination. *IEEE Transactions on Multimedia*, *26*, 4061–4072.

Zhang, Z., Wang, Z., Zhuang, S., & Wang, J. (2023). Toward action recognition and assessment using SFAGCN and combinative regression model of spatiotemporal features. *Applied Intelligence: The International Journal of Artificial Intelligence, Neural Networks, and Complex Problem-Solving Technologies*, *53*(1), 757–768.

Zhang, Y., Xiong, W., & Mi, S. (2022). Learning time-aware features for action quality assessment. *Pattern Recognition Letters*, *158*, 104–110.

Zhang, D., Zhou, D., & Liu, H. (2023). Action quality assessment for ASD behaviour evaluation. In *Proceedings - international conference on machine learning and cybernetics* (pp. 483–488).

Zhou, K., Cai, R., Ma, Y., Tan, Q., Wang, X., Li, J., et al. (2023). A video-based augmented reality system for human-in-the-loop muscle strength assessment of juvenile dermatomyositis. *IEEE Transactions on Visualization and Computer Graphics*, *29*(5), 2456–2466.

Zhou, C., Feng, D., Chen, S., Ban, N., & Pan, J. (2023). Portable vision-based gait assessment for post-stroke rehabilitation using an attention-based lightweight CNN. *Expert Systems with Applications*, Article 122074.

Zhou, K., Ma, Y., Shum, H. P. H., & Liang, X. (2023). Hierarchical graph convolutional networks for action quality assessment. *IEEE Transactions on Circuits and Systems for Video Technology*, *33*(12), 7749–7763.

Zhu, Y., Lu, W., Gan, W., & Hou, W. (2021). A contactless method to measure real-time finger motion using depth-based pose estimation. *Computers in Biology and Medicine*, *131*, Article 104282.

Zolfaghari, M., Oliveira, G. L., Sedaghat, N., & Brox, T. (2017). Chained multi-stream networks exploiting pose, motion, and appearance for action classification and detection. In *Proceedings of the IEEE international conference on computer vision* (pp. 2904–2913).