# A Survey of Video-based Action Quality Assessment

Shunli Wang[1], Dingkang Yang[1], Peng Zhai[1], Qing Yu[2], Tao Suo[2], Zhan Sun[2], Ka Li[2], Lihua Zhang[3,1]*

*Institute of AI & Robotics, Fudan University* [1], Shanghai, China
*ZhongShan Hospital* [2], Shanghai, China    *Ji Hua Laboratory* [3], Foshan, China
{slwang19, dkyang20, pzhai18, lihuazhang}@fudan.edu.cn
{ yu.qing, suo.tao, sun.zhan, li.ka }@zs-hospital.sh.cn

*Abstract*—Human action recognition and analysis have great demand and important application significance in video surveillance, video retrieval, and human-computer interaction. The task of human action quality evaluation requires the intelligent system to automatically and objectively evaluate the action completed by the human. The action quality assessment model can reduce the human and material resources spent in action evaluation and reduce subjectivity. In this paper, we provide a comprehensive survey of existing papers on video-based action quality assessment. Different from human action recognition, the application scenario of action quality assessment is relatively narrow. Most of the existing work focuses on sports and medical care. We first introduce the definition and challenges of human action quality assessment. Then we present the existing datasets and evaluation metrics. In addition, we summarized the methods of sports and medical care according to the model categories and publishing institutions according to the characteristics of the two fields. At the end, combined with recent work, the promising development direction in action quality assessment is discussed.

*Keywords*—action quality assessment, human behavior analysis, deep learning, computer vision

## I. INTRODUCTION

With the prosperity and development of video processing technology and deep neural networks (DNNs), many studies have focused on human behaviour modelling, especially in human action recognition and analysis. Existing models [1]–[4] have been able to complete complex tasks such as action classification of short videos, temporal action segmentation, classification of long-term videos and spatial-temporal action location. These studies have also been widely used in video retrieval, anomaly detection, and entertainment. However, these works can only classify and locate human action at a coarse granularity level and can not objectively evaluate the quality and compliance of some specific actions.

Based on the research of human action recognition, some studies have begun to pay attention to the action quality assessment (AQA) model. The AQA task aims to design a system that can automatically and objectively evaluate some specific actions completed by people through input videos. AQA has many practical application scenarios, such as surgical skill rating and medical rehabilitation test in medical care,

athlete posture correction and coaching system in sports, operation compliance analysis and dangerous action monitoring in industrial production. High-performance AQA systems can significantly improve the professional level of people, enhance training efficiency and reduce training costs. Models in HAR tasks only need to capture the external differences between various action categories, while models in AQA tasks focus on distinguishing internal differences within a specific action. This characteristic makes the AQA task more complex than the HAR task and puts forward higher requirements for the perception ability of the model.

Nowadays, the research on AQA tasks is still in its infancy. The action recognition task can be easily extended to rich daily behaviors, while the AQA task only applies to specific professional actions. Therefore, it is easy to find that AQA has fewer application scenarios than action recognition. Considering the data accessibility and task complexity, most studies focus on the action quality evaluation in sports and medical care. AQA methods discussed in this paper are also in these two fields. Noted that due to space and topic limitation, we only focus on the video-based action assessment methods in this paper. Methods based on kinematic sensors are excluded.

This paper provides a systematic review of the progress of video-based human action quality assessment tasks for the first time. There are three major contributions of this paper:

- The existing AQA work is comprehensively reviewed and summarized, including the critical work and progress in the field of AQA.
- This paper focuses on reviewing AQA work in sports and medical care and summarizes the existing datasets and models.
- We elaborate challenges and future opportunities of AQA to facilitate future research.

The rest of the survey is organized as follows. We first present the basic definition and form of AQA task in section II and then summarize the problems and challenges faced in AQA in section III. Then we describe existing AQA datasets and metrics in section IV. In section V, we reviewed the current AQA work from two aspects: sports and medical care. Finally, in section VI, we provide discussions and future research opportunities of AQA.

Fig. 1. Complexity of medical process: thoracentesis.
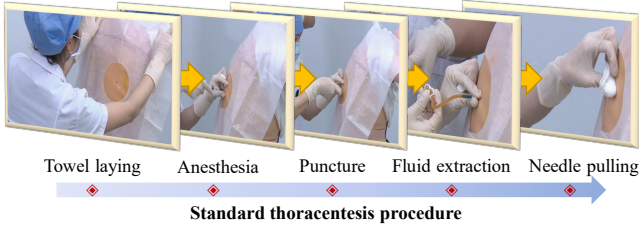


Fig. 2. The complexity of athletes' postures.

## II. BASIC FORM OF AQA

Video-based AQA is a kind of action analysis task, which automatically completes the objective evaluation of specific human action quality through videos. The AQA model and human action recognition model have some similarities. They both complete the feature extraction first and then complete various specific tasks through the network head. The general forms of AQA tasks could be classified into three types: regression scoring, grading and pairwise sorting. For clarity, we will formalize the evaluation process as follow.

Given a set of video sequences $\mathbf{S} = \{S_i\}, (i = 1, 2, \cdots, N)$ and each video can be represented as $S_i = \{f_j\}, (j = 1, 2, \cdots, T)$, where the video set contains $N$ videos, video $S_i$ contains T frames and $f_j$ denotes the $j$-th frame in $S_i$. The video will be sent to the feature extraction module and the action evaluation module, respectively.

In the *feature extraction* phase, the traditional methods first complete the extraction and selection of spatio-temporal keypoints and then adopt Discrete Fourier Transform (DFT), Discrete Cosine Transform (DCT) or linear combination to complete feature aggregation.

However, deep learning methods usually adopt deep convolution networks (DCNs) and recurrent neural networks (RNNs) to complete video embedding. The feature extraction process can be denoted as:

$$X = \mathcal{F}_{DL}(S_i), X \in \mathbb{R}^K \tag{1}$$

where $\mathcal{F}_{DL}(\cdot)$ denotes the deep neural network, such as I3D, C3D and other video feature extraction networks. $K$ represents the dimension of video features.

In the *evaluation* phase, the obtained features are sent to the evaluation module of AQA. The evaluation module can be classified into the following three types:

**Regression scoring**. The form of regression scoring usually appears in sports. Since the referees give the ground-truth scores of videos in AQA dataset in sports events, the dataset can be constructed through the broadcast videos of large-scale sports events, which is less challenging to obtain. Support Vector Regression (SVR) model or Fully Connected Network (FCN) are adopted to complete score prediction directly. Mean Square Error (MSE) is usually used as the metric in regression scoring:

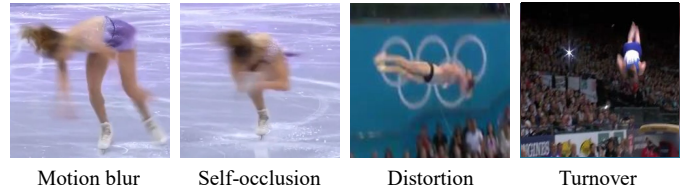$$MSE = \frac{1}{N} \sum_{i=1}^{N} (\hat{s}_i - s_i)^2 \tag{2}$$

where $\hat{s}_i$ denotes the prediction score of the AQA model on video $S_i$, and $s_i$ represents the ground-truth score of video $S_i$.

**Grading**: The form of grading usually appears in medical skill operation evaluation tasks. The operator's action will be divided into specific levels, such as *novice*, *medium* and *expert*. Therefore, the action quality assessment is transformed into a classification problem. Given dataset $\mathbf{D} = (S_i, C_i)$, $(C_i \in 1, 2, \cdots, M)$, where $C_i$ is the skill level label corresponding to video $S_i$ and there are $M$ skill levels in total. Classification accuracy is usually adopted as the metric.

**Pairwise sorting**. The pairwise sorting task takes any two videos from the video library to evaluate the action quality. Suppose that the video dataset contains $N$ videos, so there are $C_N^2$ combinations in total. The label matrix $M$ can be constructed according to the following rule:

$$M(i, j) = \begin{cases} 1, & Q(S_i) > Q(S_j) \\ -1, & Q(S_i) < Q(S_j) \\ 0, & Q(S_i) = Q(S_j) \end{cases} \tag{3}$$

where $Q(S)$ denotes the action quality assessment result of video $S$. And the label matrix $M$ satisfies the following constraints:

$$M(i, j) = -M(j, i) \tag{4}$$

The model performance is measured by pairwise sorting accuracy.

## III. CHALLENGES

The challenges faced by AQA are highly related to its application background. For example, the AQA system in sports focuses on the compliance of athletes' whole body and limb movements. AQA system in medical care focuses on doctors' gestures, medical device interaction and task completions. AQA system in industrial scenario focuses on the compliance and risk of workers' operation. At the same time, all these AQA models also face some common challenges. Therefore, the challenges faced by the AQA model can be classified into the following two categories:

**Challenges in specific areas**. The task forms in different fields are quite different, and the problems they face also have their characteristics. The action in medical care has the characteristics of high temporal complexity, semantic richness and low fault tolerance, which puts forward high requirements for the semantic understanding ability of the AQA system. For example, the standard thoracentesis shown in Fig. 1 consists of a series of sub-actions: towel laying, anesthesia, puncture,

TABLE I
SUMMARY OF EXISTING AQA DATASETS.

| Scene | Dataset | Publishing | Action class | Size | Label form | Website |
|-------|---------|-----------|:---:|:---:|-----------|---------|
| Sports | MIT-Diving [5] | ECCV 2014 | 1 | 159 | AQA score | https://www.csee.umbc.edu/ hpirsiav/quality.html |
| | MIT-Skating [5] | ECCV 2014 | 1 | 159 | AQA score | https://www.csee.umbc.edu/ hpirsiav/quality.html |
| | UNLV Dive [6] | CVPRW 2017 | 1 | 370 | AQA score | http://rtis.oit.unlv.edu/datasets.html |
| | UNLV Vault [6] | CVPRW 2017 | 1 | 176 | AQA score | http://rtis.oit.unlv.edu/datasets.html |
| | BPAD [7] | ICCV 2017 | 1 | 48 | AQA score | —— |
| | AQA-7 [8] | WACV 2019 | 7 | 1189 | AQA score | http://rtis.oit.unlv.edu/datasets.html |
| | MTL-AQA [9] | WACV 2019 | 16 | 1412 | AQA score Action class Commentary | https://github.com/ParitoshParmar/MTL-AQA |
| | FisV-5 [10] | TCSVT 2020 | 1 | 500 | TES, PCS | https://github.com/loadder/MS_LSTM.git |
| | FR-FS [11] | ACM MM 2021 | 1 | 417 | Action class | https://github.com/Shunli-Wang/TSA-Net |
| Medical Care | JIGSAWS [12] | MICCAI 2014 | 3 | 103 | AQA score | http://cirl.lcsr.jhu.edu/jigsaws. |
| Daily Life | EPIC-Skills [13] | CVPR 2018 | 3 | 196 | Pair-Rank | https://drive.google.com/file/d 1oX0dPM5IP638nB0YHt4L70aigIdqqpYr/ |
| | BEST [14] | CVPR 2019 | 5 | 500 | Pair-Rank | https://github.com/hazeld/rank-aware-attention-network |
| | Infant Grasp [15] | arXiv 2019 | 1 | 94 | Pair-Rank | —— |

fluid extraction, needle pulling, etc. However, AQA tasks in the field of sports mainly face problems such as body distortion and motion blur, as shown in Fig. 2.

**Common challenges**. In addition, there are some common challenges in various fields, such as efficiency, view occlusion, model interpretability, etc. In medical scenarios, the AQA system needs to give evaluation results and feedback in a real-time manner as much as possible to facilitate the tested personnel to correct unqualified actions in time. In sports scenarios, self-occlusion during complex movements will seriously affect human body detection and pose estimation. And the same to medical operations.

There is still a long way to go between the existing models and those with high precision, high stability, strong anti-interference and excellent interpretability. Therefore, the video-based action quality assessment system has a considerable exploration space and research significance.

## IV. DATASETS AND METRICS

We review the video-based AQA datasets in sports and medical care and briefly sort out other datasets in daily life. Then we summarize the evaluation metrics commonly used in AQA tasks. All AQA datasets are listed in Tab. II.

### A. AQA Datasets in Sports

**MIT-Diving** & **MIT-Skiing** [5]. The MIT-Diving dataset contains 159 Olympic diving competition videos. The videos are played back in slow motion with a frame rate of 60fps. The whole dataset includes 25000 frames. Each video contains 150 frames on average, with a score range of 20 to 100. Researchers visited the MIT diving team to consult the adjustment direction of each body part for evaluating the feedback system. The MIT-Skiing dataset contains 159 Olympic figure skating videos with a frame rate of 24fps. The whole dataset includes 630000 frames. Each video contains 4200 frames on average, with a score range of 0 to 100.

**UNLV Dive** & **UNLV Vault** [6]. The UNLV Dive dataset is expanded from MIT-Diving, which contains 370 videos. The UNLV Vault dataset contains 176 gymnastics videos, with an average length of about 75 frames and a score range of 0 to 20.

**Basketball Performance Assessment Dataset (BPAD)** [7]. The BPAD contains 48 videos in total. The first 24 videos were taken on the first day for training, and the last 24 videos were taken on the second day for testing. All videos are recorded from the first-person perspective. Tag label information is obtained by pairwise sorting. After professional basketball players sort the videos in pairs, 250 pairs of training and 250 pairs of tests are finally generated.

**AQA-7** [8]. The AQA-7 dataset contains seven sports and 1189 videos in total. The items and corresponding numbers are *diving*-370, *gymnastics*-176, *snowing board*-175, *skiing*-206, *synchronous diving*-88 and *trampoline*-83. The *trampoline* category is usually excluded due to the long duration. Among all videos, 803 videos are used for training and 303 for testing.

**MTL-AQA** [9]. The MTL-AQA dataset is currently the largest AQA dataset. This dataset 16 contains different events and 1412 diving samples, including single or synchronous diving, male or female athletes, 3m springboard or 10m platform. In addition to providing ground-truth scores for each video, this dataset also provides fine-grained action classes and comment information, so it can be used to complete AQA, action recognition and comment generation. The scores and difficulty degrees of all videos are given by 7 judges. Among all videos, 1059 videos are used for training and 353 for testing.

**FisV-5** [10]. The FisV-5 dataset contains 500 figure skating competition videos, with an average duration of 2 minutes and 50 seconds. The score label of the video is given by 9 professional judges and is divided into two parts: Total Element Score (TES) and Total Program Component Score (PCS). Among all videos, 400 videos are used for training

and 100 for testing.

**Fall Recognition in Figure Skating (FR-FS)** [11]. Existing AQA datasets of figure skating [5], [10] only contain long videos, which will lead to the inundation of detailed information, and finally affect the evaluation performance of the AQA model. To solve this problem, Wang *et al.* [11] proposed FR-FS to recognize figure skating falls and planned to build a delicate granularity AQA system gradually. Videos in FR-FS contains the movements of the athlete's take-off, rotation, and landing. The FR-FS dataset contains 417 videos. Among these videos, 276 are smooth landing videos, and 141 are fall videos.

### B. AQA Datasets in Medical Care

Compared with sports, video-based AQA datasets are more scarce in medical care. Each research institution builds its data collection platform to construct the AQA datasets, so there is very little open source-work. Currently, the only open-source work in medical care is JIGSAWS.

**JIGSAWS** [12]. The full name of JIGSAWS is JHU-ISI Posture and Skill Assessment Working Set. JIGSAWS includes three surgical behaviors: *suturing*, *needle passing* and *knot tying*. Each video is scored in many aspects, such as operation and result quality. The final score is the summation of sub-scores. All videos in the dataset are recorded with stereo cameras. Eight doctors with different skills are invited as operators. It contains 36 trials of knot tying, 28 trials of need passing and 39 trials of suturing.

### C. Other AQA Datasets

In addition to the datasets in sports and medical care, some researchers have explored tasks in daily life and constructed AQA datasets. Since there is no strict operation process regulation for these actions, the evaluation metric of pair-wise ranking is adopted.

**Epic skills 2018** [13]. The Epic skills 2018 dataset consists of three sub-datasets: 1. *Dough-Rolling* is collected from the CMU-MMAC dataset [16] which contains actions performed 33 different participants. 2. *Drawing*: Four participants were asked to complete painting imitations five times. Two reference images are provided: a cartoon of Sonic the Hedgehog and a gray scale photo of a hand. 3. *Chopstick-Using*: Eight participants were asked to use chopsticks to carry the beans in the lunch box for five times. The action quality is determined by the successful transport times within one minute.

**BEST** [14]. The full name of BEST is Bristol Everyday Skill Tasks. This dataset is constructed for skill evaluation of long-term video with an average duration of 188s. It contains 500 videos in total and 5 daily tasks: *scrambling eggs*, *braiding hair*, *tying a tie*, *making an origami crane* and *applying eyeliner*.

**Infinite grasp dataset** [15]. This dataset is introduced from psychological research and contains 94 videos of infants grabbing objects. The length of the video varies from 80 to 500. Five psychologists were asked to label the videos in pairs, resulting in 4371 pairs, of which 3318 pairs with pairwise differences.
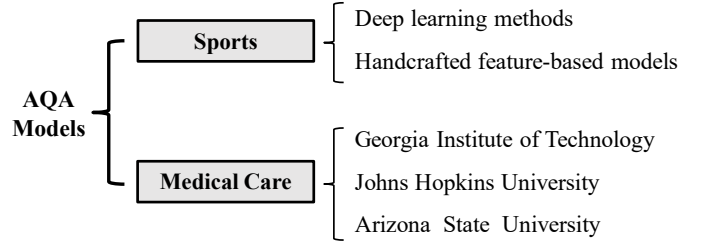


Fig. 3. Categorization methodology adopted by this paper.

### D. Performance Metrics

Due to the rich task forms, there is no general evaluation metric for AQA tasks. In section 2, the AQA models are categorized into the following three types: regression scoring, grading and pairwise sorting. Therefore, the corresponding evaluation metrics are mean square error, classification accuracy and Spearman correlation coefficient $\rho$. The Spearman correlation coefficient is defined as follow.

$$\rho = \frac{\sum_i (p_i - \bar{p})(q_i - \bar{q})}{\sqrt{\sum_i (p_i - \bar{p})^2 \sum_i (p_i - \bar{q})^2}} \tag{5}$$

## V. AQA MODELS

From the perspective of development history, the emergence of AQA models in medical care is earlier than in sports because medical skills have two characteristics. On the one hand, experienced doctors have to evaluate the operation skills of each novice in the standard resident training process, and the workload is enormous. On the other hand, with the continuous development of medical robots (such as *Da Vinci Robot*), robots can replace doctors to complete some medical operations to a certain extent. It is necessary to specify a recognized standard to evaluate robot operations. Before the prosperity of deep learning, these factors have aroused people's interest in skills assessment in medical care. Generally speaking, most medical skill evaluation methods use traditional features to model video information. We will describe these efforts in detail below.

Compared with medical care, researches on action quality assessment on sports video started relatively late. These studies mainly focus on the actions of athletes in the Olympic Games, such as diving, gymnastics, trampoline and other sports mentioned above. The rise of deep learning in recent years provides a powerful tool for sports video assessment. Most studies have achieved considerable results through CNN and RNN. This is also the development trend of future research. Considering the above internal differences between sports and medical care, we still take the research topic of the paper as a taxonomic basis. Fig. 3 shows the categorization methodology adopted by this paper.

### A. AQA Models in Sports

In this subsection, we focus on the action quality assessment methods in sports video. We divide these methods into two categories: models based on deep learning and handcrafted

TABLE II
SUMMARY OF EXISTING AQA METHODS.

| Ref. | Published | Dataset | Backbone | Network Head | Metrics | Website |
|---|---|---|---|---|---|---|
| [17] | RO-MAN 2016 | Powerlifting | RSO Net | Feedback | Time Series Analysis | —— |
| [18] | ICIP 2018 | AQA-7 JIGSAWS | P3D | Reg | Spearman Rank Cor. | https://github.com/YeTianJHU/diving-score |
| [19] | PCM 2018 | UNLV-Diving UNLV-Vault MIT-Skate | C3D | Rank loss Reg | Spearman Rank Cor. Mean Euclidean Dis. | —— |
| [8] | WACV 2019 | AQA-7 | C3D | Reg | Spearman Rank Cor. | http://rtis.oit.unlv.edu/datasets.html |
| [9] | CVPR 2019 | MTL-AQA | C3D | Caption Cls Reg | Spearman Rank Cor. | https://github.com/ParitoshParmar/MTL-AQA |
| [20] | ICCV 2019 | AQA-7 JIGSAWS | I3D | Reg | Spearman Rank Cor. | —— |
| [10] | TCSVT 2020 | FisV MIT-Skate | C3D | Reg | Spearman Rank Cor. Mean Square Error | https://github.com/loadder/MS_LSTM.git |
| [21] | CVPR 2020 | AQA-7 MTL-AQA JIGSAWS | I3D | Reg | Spearman Rank Cor. | https://github.com/nzl-thu/MUSDL |
| [11] | ACM MM 2021 | FR-FS AQA-7 MTL-AQA | I3D | Reg Cls | Spearman Rank Cor. | https://github.com/Shunli-Wang/TSA-Net |
| [7] | ICCV 2017 | First-Person Basketball | Att. to scale | Pair-Rank | Ranking Acc. | —— |
| [13] | CVPR 2018 | EPIC-Skills JIGSAWS | TSN | Pair-Rank | Ranking Acc. | —— |
| [14] | CVPR 2019 | BEST EPIC-Skills | I3D | Pair-Rank | Ranking Acc. | https://github.com/hazeld/ rank-aware-attention-network |
| [15] | ICCVW 2019 | Infant Grasp EPIC-Skills JIGSAWS | ResNet101 | Pair-Rank | Ranking Acc. | —— |

features. Note that since the models tested on the daily life datasets [13]–[15] are all based on deep learning, we divide them into sports, even if these methods are not oriented to professional sports.

*1) Models based on deep learning:* Deep learning methods usually adopt 2D-CNN, 3D-CNN and LSTM to extract and aggregate the features of videos and then complete the assessment task. As shown in Fig.4, these deep learning-based AQA frameworks are composed of a video feature extraction module and an evaluation module. The form of the evaluation module is highly related to the type of assessment task. According to the framework position concerned by these works, these methods can be categorised into the following two classes: *structure designing* and *loss designing*.

Some studies focus on the design of the network structure to extract more discriminative features. In order to learn the local and global sequential information in long-term videos, Xu *et al.* [10] proposed Self-Attentive LSTM and Multi-scale Convolutional Skip LSTM to predict TES and PCS in figure skating. Xiang *et al.* [18] divided the diving process into four stages: beginning, jumping, dropping and entering into the water. They adopted four independent P3D [22] models to complete feature extraction. Pan *et al.* [20] proposed a graph-based joint relation model to analyze the motion of human nodes. Human nodes and their correlation are modelled by the proposed Joint Commonality Module and the Joint Difference Module. Parisi *et al.* [17] proposed a new recurrent neural

network, which adopts the growing self-organizing structure to learn the body motion sequence and complete the matching. Kim *et al.* [23] modeled the action as a structured process. Action units are encoded from dense trajectories with a LSTM network. Wang *et al.* [11] proposed a feature aggregation mechanism named tube self-attention module to efficiently generate rich spatio-temporal contextual information by adopting sparse feature interactions.

Some studies focus on the design of network loss to obtain better action quality assessment performance. Li *et al.* [19] proposed an end-to-end framework that takes C3D as the feature extractor and designs a ranking loss integrated with the MSE loss. Parmar *et al.* [9] explored the AQA model in the multi-task learning scenario. They proposed three parallel prediction tasks: action recognition, comment generation, and AQA score regression, and then verified the effectiveness by the proposed C3D-AVG-MTL framework. Tang *et al.* [21] proposed an uncertainty-aware score distribution learning approach. They incorporated the difficulty degree into the modeling process and simulated the scoring process more realistically.

In addition, some studies focus on the quality comparison task of paired actions. Bertasius *et al.* [7] proposed a model based on the first perspective videos in basketball games. This model adopted a convolutional-LSTM network to detect events and evaluate the quality of any two movements. Doughty *et al.* [13], [14] proposed the task of sorting paired videos. These
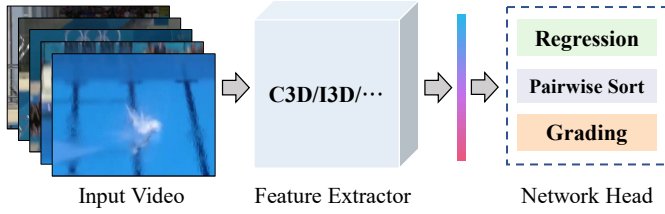
Fig. 4. Framework of AQA model based on deep learning.



**GIT**
Simulated surgery

**JHU**
Da Vinci Robot

**ASU**
Laparoscopic surgery

Fig. 5. Dataset acquisition platforms of three three research series.

two works proposed EPIC-Skills and BEST, respectively. In [13], they took temporal segment network (TSN) [24] as the backbone and proposed a ranking loss based on margin loss. [14] is the subsequent work of [13], which released the BEST dataset with richer scene information and proposed a model with a more complex sorting loss function based on I3D [3]. Li *et al.* [15] proposed a spatial attention model based on RNN. Attention pooling and temporal aggregation mechanisms were adopted to complete feature extraction. This model was tested on the released Infant Grasp dataset.

*2) Models based on handcrafted features:* Traditional AQA methods take handcrafted features as discriminant features. These feature extraction methods mainly appeared before 2014. In [25], Gordon first proposed video-based AQA tasks and took the gymnastics scoring as an example to analyse the realizability of AQA technology. Ilg *et al.* [26] established correspondence at the global action sequence level and the single motion element level. A spatio-temporal deformation model is proposed for example comparison. Çeliktutan *et al.* [27] proposed a dynamic skeleton sequence alignment method based on graph and applied it to action recognition and assessment. Pirsiavash *et al.* [5] proposed an analysis method in the frequency domain to complete the quality evaluation based on the combination of low-level visual features and high-level pose features. Paiement *et al.* [28] proposed a general method for on line assessment of motion quality based on Kinect data. After reducing the dimension of noisy data by robust nonlinear manifold learning, the matching degree between observation data and prediction data is calculated frame by frame based on the Markov hypothesis.

### B. AQA Models in Medical Care

Some institutions have made a series of contributions to medical skills assessment, such as Johns Hopkins University (JHU), Georgia Institute of Technology (GIT), Arizona State University (ASU), etc. Interestingly, compared with the action quality assessment of sports video, it is easy to find that these works have a strong inheritance. Therefore, we will creatively review the literature according to the publishing organization as the classification basis in this part.

Due to the strong professionalism of medical skill assessment, there is no unified dataset as a benchmark. The series of studies of the institutions mentioned above only focus on a specific scenario. Studies of GIT focus on the medical behavior quality assessment under the Objective Structured Assessment of Technical Skill (OSATS) system [29]. Studies of GIT focus

on the action quality assessment of the surgical robots. Studies of ASU focus on the action quality assessment using surgical training equipment. Fig. 5 shows the scenarios concerned by these three series of studies. Next, we will introduce these three types of studies in chronological order respectively.

*1) GIT:* The series studies of Georgia Institute of Technology focus on the Surgical skills evaluation under the OSATS system [29]. Inspired by LDS and bag-of-features (BoF) methods [30], [31], Sharma *et al.* [32] proposed an enhanced BoW model to solve the problem that traditional models cannot process implicit temporal information and causality. In [32], [33], Sharma *et al.* evaluated medical skills through video motion texture and sequential motion texture modeling respectively. Based on [32], Sharma *et al.* [33] introduced a data-driven temporal window to complete the modeling of sequential motion features. Zia *et al.* [34], [35] adopted the same methods, and the latter is the extended journal version of the former. In [34], after the extraction of STIP keypoints, K-means clustering is adopted to generate time series, and then DCT and DFT are utilized to explore the temporal characteristics. To solve the problem of deficient information in traditional methods, Zia *et al.* [36] proposed an off-line unsupervised temporal clustering method and divided the long-term surgical operation into small steps, which is similar to those studies of JHU. Zia *et al.* [37] created a surgical skill evaluation dataset based on camera and accelerometer and proposed a model that combines approximate entropy and cross approximate entropy to quantify the predictability and regular beat of time series.

*2) JHU:* The series of studies of Johns Hopkins University focused on the scene of robotic minimally invasive surgery (RMIS). Most of the studies take *Da Vinci Robot* as the equipment for data acquisition. Haro *et al.* [30] and Zappella *et al.* [31] explored the automatic classification of gestures and skills in robot assisted surgery. Haro *et al.* [30] proposed a surgical posture classification method based on linear dynamic system (LDS) and bag-of-features and carried out more experiments in the extended journal version [31]. Although these two studies did not directly assess the operation quality, they laid a solid foundation for follow-up work. Malpani *et al.* [38] proposed a model that automatically scores surgical segments and completes the quality ranking by pairwise comparison. Gao *et al.* [12] constructed and released the JIGSAWS dataset and expounded the concept of *JHU Language of Surgery* project in detail. Some early studies complete the evaluation

of medical skills only based on kinematic data. For example, robot operation assessment based on HMM [39], temporal modeling of kinematic data based on RNN [40], endoscopic sinus surgery pathway evaluation [41], and skill evaluation in septoplasty [42].

*3) ASU:* The series of studies of Arizona State University mainly focus on the fundamentals of laparoscopic surgery (FLS-Box). Traditional medical skill evaluation methods usually need contact sensors which will affect the operation of doctors. To overcome this disadvantage, [43], [44] proposed a surgical skill method based on visual information. Based on Islam *et al.* [43], [44] improved the algorithm to real-time, added feedback function and tested the model in the real surgical video. In order to deal with the semantic gap between low-level visual features and high-level surgical skills, Chen *et al.* [45] fused multi-modal information and completed meaningful surgical skill analysis by looking for hidden space. Zhang *et al.* [46]–[48] adopted HMM and its variants to solve AQA problems and made a series of progress. In [46], they proposed a hierarchical Dirichlet process Markov model for the classification of surgical skills. In [47], they proposed a relative HMM model, which only takes the relative ranking between paired videos as input for training. Based on [47], researchers in [48] improved the model and extended the application scenario to a language-based emotion recognition task. Based on the previous work, Chen *et al.* [49] further integrated and improved the function of the action quality assessment system and improved the interactivity.

*C. Summary of AQA models*

We have reviewed the work related to AQA in sports and medical care. Through the above classification and summary, we can get the following findings.

- From the perspective of development history, the AQA problem in medical care was put forward earlier than that in sports. This is because the system that can automatically and objectively evaluate surgical skills has more urgent needs and greater practical significance.
- From the perspective of feature extractor, most of the methods before 2014 constructed the AQA system through handcrafted features and traditional classifiers. In recent years, some general network backbones have been proposed [50]–[53]. More and more work began to use DNNs to build AQA model due to the stronger representation ability and achieved excellent performance.
- From the perspective of the datasets, it is more challenging to construct datasets in medical care than in sports. Datasets in sports usually consist of Olympic Games, while datasets in medical care are more professional and have more institutional background and continuity.

## VI. PROSPECT OF FUTURE WORK

Combined with recent progress in human action understanding, this section looks forward to the future research direction of AQA task from two aspects: dataset and model, as shown in Fig.6. At the dataset level, existing AQA datasets are small
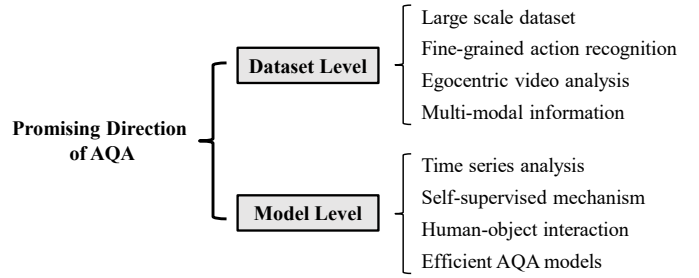


Fig. 6. Some promising research directions of AQA task in future.

in scale and have weak semantic richness. We look forward to the proposal of datasets with larger scale and richer semantic information in the future. In addition, some novel datasets and models have been proposed recently in HAR tasks. These studies focus on fine-grained action recognition [50], [51], egocentric action recognition [52], [53] and multimodal information [54], [55]. New scenarios and datasets can promote the exploration and development of AQA.

At the model level, there are still many efforts to be paid for efficient and accurate AQA models. For example, how to make full use of temporal information to model the action, use unsupervised methods to avoid expert annotation, take human-object interaction into account, and design efficient models. Recent studies in HAR has made a preliminary exploration on these problems, such as effective temporal modeling [56], [57], self-supervised mechanism [58], [59], taking human-object interaction information into consideration [60], [61], and efficient designing of HAR models [62], [63]. These methods provide a solid foundation for the follow-up development of AQA.

Most existing AQA methods assume that the operation sequence is known and the task is relatively easy. There is no work to explore the long-term complex action quality assessment. However, some studies have explored unsupervised structural modelling of complex actions, such as instruction video analysis [64], [65] and unsupervised process self discovery [66]. These works explore the task of video analysis under more stringent conditions and put forward higher and requirements for AQA models.

## VII. CONCLUSION

Video-based action quality assessment is a critical and challenging problem in computer vision and has received considerable attention. As the first comprehensive survey of AQA tasks, this paper summarizes the AQA datasets and methods according to the categorization methodology of sports and medical care. The challenges of the AQA task in two fields are discussed in detail. In the end, we discuss the future development direction combined with recent studies. Despite the considerable performance, existing AQA methods can only analyse the video under ideal conditions and can not handle complex scenes. We believe that high-performance AQA systems will bring great convenience to our life, especially in sports, medical care and skill training.

REFERENCES

[1] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2014.

[2] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[3] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[4] D. Tran, L. D. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015.

[5] H. Pirsiavash, A. Torralba, and C. M. Vondrick, "Assessing the quality of actions," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2014.

[6] P. Parmar and B. T. Morris, "Learning to score olympic events," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017.

[7] G. Bertasius, H. S. Park, S. X. Yu, and J. Shi, "Am i a baller? basketball performance assessment from first-person videos," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019.

[8] P. Parmar and B. T. Morris, "Action quality assessment across multiple actions," in *Winter Conference on Applications of Computer Vision (WACV)*, 2018.

[9] ——, "What and how well you performed? a multitask learning approach to action quality assessment," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[10] C. Xu, Y. Fu, B. Zhang, Z. Chen, and X. Xue, "Learning to score figure skating sport videos," *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, vol. PP, no. 99, pp. 1–1, 2019.

[11] S. Wang, Y. D., Z. P., C. C., and Z. L., "Tsa-net: Tube self-attention network for action quality assessment," in *Proceedings of ACM international conference on Multimedia (ACM MM)*, 2021.

[12] G. Y, V. S. S, R. C. E, and et al., "Jhu-isi gesture and skill assessment working set (jigsaws): A surgical activity dataset for human motion modeling," in *Modeling and Monitoring of Computer Assisted Interventions (M2CAI)*, 2014.

[13] H. Doughty, D. Damen, and W. Mayol-Cuevas, "Who's better? who's best? pairwise deep ranking for skill determination," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[14] H. Doughty, W. Mayol-Cuevas, and D. Damen, "The pros and cons: Rank-aware temporal attention for skill determination in long videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[15] Z. Li, Y. Huang, M. Cai, and Y. Sato, "Manipulation-skill assessment from videos with spatial attention network," in *Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCVW)*, 2019.

[16] F. D. la Torre Frade, J. K. Hodgins, A. W. Bargteil, X. M. Artal, J. C. Macey, A. C. I. Castells, and J. Beltran, "Guide to the carnegie mellon university multimodal activity (cmu-mmac) database," Carnegie Mellon University, Tech. Rep., 2008.

[17] G. I. Parisi, S. Magg, and S. Wermter, "Human motion assessment in real time using recurrent self-organization," in *IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 2016.

[18] X. Xiang, Y. Tian, A. Reiter, G. D. Hager, and T. D. Tran, "S3d: Stacking segmental p3d for action quality assessment," in *Proceedings of the International Conference on Image Processing (ICIP)*, 2018, pp. 928–932.

[19] Y. Li, X. Chai, and X. Chen, "End-to-end learning for action quality assessment," in *Advances in Multimedia Information Processing – PCM*, 2018.

[20] J. H. Pan, J. Gao, and W. S. Zheng, "Action assessment by joint relation graphs," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019.

[21] Y. Tang, Z. Ni, J. Zhou, D. Zhang, and J. Zhou, "Uncertainty-aware score distribution learning for action quality assessment," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[22] Z. Qiu, T. Yao, and T. Mei, "Learning spatio-temporal representation with pseudo-3d residual networks," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[23] S. T. Kim and M. R. Yong, "Evaluationnet: Can human skill be evaluated by deep networks?" *arXiv:1705.11077*, 2017.

[24] L. Wang, Y. Xiong, W. Zhe, Q. Yu, and L. V. Gool, "Temporal segment networks for action recognition in videos," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. PP, no. 99, 2017.

[25] A. S. Gordon, "Automated video assessment of human performance," *Proceedings of AI-ED 95*, 1995.

[26] W. Ilg, J. Mezger, and M. Giese, "Estimation of skill levels in sports based on hierarchical spatio-temporal correspondences," in *Pattern Recognition*, 2003.

[27] O. Çeliktutan, C. B. Akgul, C. Wolf, and B. Sankur, "Graph-based analysis of physical exercise actions," in *ACM Int. Workshop on Multimedia Indexing and Information Retrieval for Healthcare*, 2013.

[28] A. Paiement, L. Tao, S. Hannuna, M. Camplani, and M. Mirmehdi, "Online quality assessment of human movement from skeleton data," in *British Machine Vision Virtual Conference (BMVC)*, 2014.

[29] J. A. Martin, G. Regehr, R. Reznick, H. Macrae, J. Murnaghan, C. Hutchison, and M. Brown, "Objective structured assessment of technical skill (osats) for surgical residents," *British Journal of Surgery*, 1997.

[30] B. B. Haro, L. Zappella, and R. Vidal, "Surgical gesture classification from video data," in *Int. Conf. on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2012.

[31] L. Zappella, B. Béjar, G. Hager, and R. Vidal, "Surgical gesture classification from video and kinematic data," *Medical image analysis*, 2013.

[32] Y. Sharma, V. Bettadapura, T. Ploetz, N. Hammerla, S. Mellor, R. McNaney, P. Olivier, S. Deshmukh, A. McCaskie, and I. Essa, "Video based assessment of osats using sequential motion textures," in *Modeling and Monitoring of Computer Assisted Interventions (M2CAI)*, 2016.

[33] Y. Sharma, T. Plötz, N. Hammerld, S. Mellor, R. Mcnaney, P. Olivier, S. Deshmukh, A. Mccaskie, and I. Essa, "Automated surgical osats prediction from videos," in *IEEE International Symposium on Biomedical Imaging (ISBI)*, 2014.

[34] A. Zia, Y. Sharma, V. Bettadapura, E. L. Sarin, and I. Essa, "Automated assessment of surgical skills using frequency analysis," in *Int. Conf. on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015.

[35] A. Zia, Y. Sharma, V. Bettadapura, E. L. Sarin, T. Ploetz, M. A. Clements, and I. A. Essa, "Automated video-based assessment of surgical skills for training and evaluation in medical schools," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 11, no. 9, pp. 1623–1636, 2016.

[36] A. Zia, C. Zhang, X. Xiong, and A. M. Jarc, "Temporal clustering of surgical activities in robot-assisted surgery," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 12, no. 7, pp. 1171–1178, 2017.

[37] A. Zia, Y. Sharma, V. Bettadapura, E. L. Sarin, and I. A. Essa, "Video and accelerometer-based motion analysis for automated surgical skills assessment," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 13, no. 3, pp. 443–455, 2018.

[38] A. Malpani, S. S. Vedula, C. C. G. Chen, and G. D. Hager, "Pairwise comparison-based objective score for automated skill assessment of segments in a surgical task," in *Information Processing in Computer-Assisted Interventions (IPCAI)*, 2014.

[39] L. Tao, E. Elhamifar, S. Khudanpur, G. D. Hager, and R. Vidal, "Sparse hidden markov models for surgical gesture classification and skill evaluation," in *Information Processing in Computer-Assisted Interventions (IPCAI)*, 2012.

[40] R. S. DiPietro, C. Lea, A. Malpani, N. Ahmidi, S. S. Vedula, G. I. Lee, M. R. Lee, and G. D. Hager, "Recognizing surgical activities with recurrent neural networks," in *Int. Conf. on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2016.

[41] N. Ahmidi, G. D. Hager, L. Ishii, G. L. Gallia, and M. Ishii, "Robotic path planning for surgeon skill evaluation in minimally-invasive sinus surgery," in *Int. Conf. on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2012.

[42] N. Ahmidi, P. Poddar, J. D. Jones, S. S. Vedula, L. Ishii, G. D. Hager, and M. Ishii, "Automated objective surgical skill assessment in the operating room from unstructured tool motion in septoplasty," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 10, no. 6, pp. 981–991, 2015.

[43] G. Islam, K. Kahol, J. Ferrara, and R. Gray, "Development of computer vision algorithm for surgical skill assessment," in *Int. Conf. on Ambient Media and Systems*, 2011.

[44] G. Islam, B. Li, and K. Kahol, "An affordable real-time assessment system for surgical skill training," in *Int. Conf. on Intelligent User Interfaces (IUI)*, 2013.

[45] L. Chen, Q. Zhang, Q. Tian, and B. Li, "Learning skill-defining latent space in video-based analysis of surgical expertise - A multi-stream fusion approach," in *Medicine Meets Virtual Reality (MMVR)*, 2013.

[46] Q. Zhang and B. Li, "Video-based motion expertise analysis in simulation-based surgical training using hierarchical dirichlet process hidden markov model," in *Int. ACM workshop on Medical Multimedia Analysis and Retrieval*, 2011.

[47] ——, "Relative hidden markov models for evaluating motion skill," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[48] ——, "Relative hidden markov models for video-based evaluation of motion skills in surgical training," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 37, no. 6, pp. 1206–1218, 2015.

[49] L. Chen, Q. Zhang, P. Zhang, and B. Li, "Instructive video retrieval for surgical skill coaching using attribute learning," in *IEEE International Conference on Multimedia and Expo (ICME)*, 2015.

[50] D. Shao, Y. Zhao, B. Dai, and D. Lin, "Finegym: A hierarchical video dataset for fine-grained action understanding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[51] D. Damen, H. Doughty, G. Farinella, S. Fidler, A. Furnari, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price, and M. Wray, "The epic-kitchens dataset: Collection, challenges and baselines," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2020.

[52] S. Sudhakaran, S. Escalera, and O. Lanz, "Lsta: Long short-term attention for egocentric action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[53] M. Liu, S. Tang, Y. Li, and J. Rehg, "Forecasting human-object interaction: Joint prediction of motor attention and actions in first person video," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2019.

[54] A. Piergiovanni, A. Angelova, and M. S. Ryoo, "Evolving losses for unsupervised video representation learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[55] H. Alwassel, D. Mahajan, L. Torresani, B. Ghanem, and D. Tran, "Self-supervised learning by cross-modal audio-video clustering," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[56] B. Zhou, A. Andonian, A. Oliva, and A. Torralba, "Temporal relational reasoning in videos," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[57] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019.

[58] S. Benaim, A. Ephrat, O. Lang, I. Mosseri, W. T. Freeman, M. Rubinstein, M. Irani, and T. Dekel, "Speednet: Learning the speediness in videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[59] J. B. Alayrac, A. Recasens, R. Schneider, R. Arandjelovic, and A. Zisserman, "Self-supervised multimodal versatile networks," *arXiv:2006.16228*, 2020.

[60] R. Goyal, S. E. Kahou, V. Michalski, J. Materzynska, S. Westphal, H. Kim, V. Haenel, I. Fruend, P. Yianilos, M. Mueller-Freitag, F. Hoppe, C. Thurau, I. Bax, and R. Memisevic, "The "something something" video database for learning and evaluating visual common sense," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[61] X. Wang, L. Zhu, Y. Wu, and Y. Yang, "Symbiotic attention for egocentric action recognition with object-centric alignment," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2020.

[62] C. Feichtenhofer, "X3d: Expanding architectures for efficient video recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[63] J. Lin, C. Gan, and S. Han, "Tsm: Temporal shift module for efficient video understanding," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019.

[64] J.-B. Alayrac, P. Bojanowski, N. Agrawal, J. Sivic, I. Laptev, and S. Lacoste-Julien, "Learning from narrated instruction videos," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 40, no. 9, pp. 2194–2208, 2018.

[65] J.-B. Alayrac, J. Sivic, I. Laptev, and S. Lacoste-Julien, "Joint discovery of object states and manipulation actions," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[66] D. Damen, T. Leelasawassuk, and W. Mayol-Cuevas, "You-do, i-learn: Egocentric unsupervised discovery of objects and their modes of interaction towards video-based guidance," *Computer Vision and Image Understanding (CVIU)*, 2016.