ACTION QUALITY ASSESSMENT FOR ASD BEHAVIOUR EVALUATION

DINGHUANG ZHANG[1], **DALIN ZHOU**[1], **HONGHAI LIU**[1]*
[1]School of Computing, University of Portsmouth, Portsmouth, P01 3HE, UK
E-MAIL: dinghuang.zhang@port.ac.uk, dalin.zhou@port.ac.uk, Honghai.liu@port.ac.uk
∗ Corresponding author.

Abstract:

Given the current increasing prevalence of autism, expensive and time-consuming manual diagnosis is highly detrimental to the management of the condition. With the development of computer-based methods of human behavioural analysis, these methods are expected to provide more accurate, objective and reproducible methods of early screening and diagnosis of autism. To advance the field of behavioural quantification in autism research, this study utilises human skeletal behavioural data from publicly available autism datasets and ADOS scores from clinical professionals in a first attempt to build deep neural networks that can predict ADOS scores from behavioural data using the AQA approach. This paper finds a moderately correlated between the ground truth ADOS score and the predicted ADOS score, it reveals the potential use of the AQA method in ASD diagnoses.

Keywords:

AQA; ASD; Behaviour Evaluation; CNN-LSTM; Skeleton;

1 Introduction

Early detection and diagnosis of autism spectrum disorder (ASD) are vital for ensuring children receive the necessary support and interventions. However, the shortage of clinical professionals and the lengthy diagnostic process based on manual behaviour observation pose significant barriers. This emphasises the urgent need for the development of efficient and effective methods to enhance the diagnosis procedure. The field of Human action evaluation (HAE) has emerged as a promising solution to this challenge. This field leverages computer vision technology providing valuable insights into behaviour patterns. Utilising Vision-based human behaviour analysis methods, the atypical behaviour patterns in individuals with ASD can be objectively measured to provide valuable insights and objective behaviour measurement to aid in the diagnosis process.

The HAE aims to design computational models that represent the dynamic processes of human movement and to develop evaluation techniques to measure the quality of completion of human actions. This is an important area of research as it helps to advance the development of human movement evaluation systems, which are important in areas such as movement scoring, clinical assessment, and skill assessment. And the goal of HAE is to develop models that can accurately evaluate human actions. Under that goal, the Action quality assessment (AQA) is the process of evaluating the quality of the completion of an action. This is a sub-field of HAE that specifically focuses on assessing the quality of an action rather than simply recognition or predicting the action. The AQA technique aims to provide a comprehensive quantitative assessment of all aspects of movement quality, such as accuracy, consistency, fluency and naturalness etc. The AQA project aims to develop an automated system that can assess particular actions performed by individuals using behaviour data inputs to achieve objective evaluations [1]. This system's ability to quantify behaviour has various potential applications, including assessing athletic prowess, gauging an individual's proficiency, or enhancing the accuracy of systems that forecast and detect human motion.

In other words, AQA models concentrate on identifying internal distinctions inside particular actions, in contrast to human action recognition models, which capture the external differences between multiple action categories. As a result, the AQA test is more difficult than the human action identification task and exerts more pressure on the model's capacity for perception [1]. The three main categories of AQA activities are regression scoring, grading, and pairwise sorting.

Compared to human action recognition, which has the ability to encompass a wide range of everyday behaviours, the scope of AQA is limited to specific professional actions, leading to fewer potential applications. Consequently, current research efforts in AQA have mainly concentrated on sports and medical care due to data availability and task complexity considerations. The central objective of AQA is to develop computational models that capture the dynamic processes of human movement and to establish evaluation techniques that measure the quality of action completion. Recent years have witnessed the proposal of various approaches towards achieving this goal.

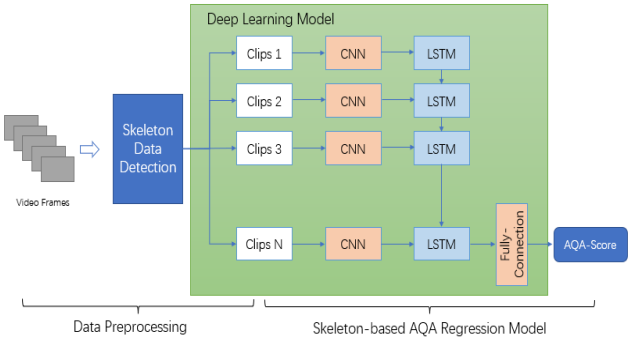

FIGURE 1. Overview of proposed AQA-based approach

AQA is primarily concerned with assessing the degree of proficiency in performing specific movements or behaviours, as reflected in the complete score. This score can be used to facilitate deeper analysis, such as physical rehabilitation, skill training, and anomaly detection. Ultimately, AQA seeks to provide objective evaluations of human action quality based on data captured from human behaviour.

AQA has attracted growing attention in recent years since it plays a crucial role in many real-world applications, including sports, healthcare and others [2, 3]. It is a data-driven regression method that evaluates how well a specific action is performed. In different to the classification and detection, the challenge of AQA is it requires the model to predict fine-grained scores from videos that describe the same action. In other words, the training objective of AQA is to create a mapping from behavioural features to scores, i.e. to learn how to score. Therefore, AQA holds the promise of integrating human observation and computer-aided systems, i.e. forming a dataset based on manually observed videos of participant behaviour and manually scored results and training AQA models.

In this chapter, an attempt is made to use machine-learning methods to aid in behavioural observation. Unlike feature engineering, deep learning models are maturely adapted to various end-to-end applications. They offer advantages in language translation, image recognition, and action recognition. Given the scale-based behavioural observations of ASD now used, establishing a direct connection between behaviour and the behaviour observation score could considerably help solve the current difficulty in diagnosing ASD.

2 Related Work

AQA involves the development of deep learning models that can objectively and automatically evaluate specific actions performed by individuals based on input videos. There are three main forms of AQA tasks: regression scoring, grading, and pairwise sorting. In the study "Learning To Score Olympic Events" [4], the authors built a regression model that could determine the score of Olympic events for a specific sport. Similarly, in the context of assessing human behaviour, we aim to build scores that can reflect clinician behavioural observations from the behaviour of individuals with ASD. High-performance AQA systems have the potential to significantly enhance the professional standards of individuals, increase training efficiency, and reduce training costs.

In theory, the regression scoring method in AQA builds a machine learning model to connect the rating scale with actions by treating the scores as a ground truth label. Therefore, the trained model can predict a scale for raw action. This means a well-trained AQA model based on the ASD behaviour dataset and corresponding behaviour observation score can release the workload for clinicians from time-consuming behaviour observation. Therefore in order to construct an AQA regression model for ASD behavioural scores, a dataset of autistic behaviours with scores is first required. To our knowledge, the only publicly available dataset that covers ASD behaviours and clinician scores is the DREAM dataset[5], which consists of participants' skeleton-based upper body movement data, gaze data, and corresponding scores rated by clinicians based on ADOS. The ADOS module is a semi-structured ASD diagnostic observation containing four areas: communication, reciprocal social interaction, play, and stereotypical behaviour and restricted interest [6].

Nowadays, AQA has attracted increasing attention in computer vision for its practical applications. The current AQA study dataset contains mainly sports scoring, such as MIT-Diving & Skiing[7], AQA-7[8], FisV-5 [9]; skill level scoring, such as JIGSAWS[10]; and daily life scoring, such as Epic skills 2018 [11], BEST[12] and Infinite grasp dataset [13]. Similar to the action recognition task, AQA uses deep neural networks for feature extraction and then selects the appropriate performance metrics for model training, depending on the type of task. In current AQA research, human motion data is usually represented by two categories, video or skeletal data.

This paper attempts to develop a learning framework for assessing ASD behaviours based on the AQA model using a dataset of ASD behaviours with corresponding clinician scores. First, we identified the use of the DREAM dataset, especially

the regression model labels and the kinematic skeletal data. Then we adopted a deep learning model and a similar learning strategy with the Olympic Events scoring to build the AQA regression model [4]. The contributions of this chapter are as follows: Introducing AQA to the study of behavioural observations in ASD suggests a new research direction for computer-assisted ASD research. We first apply the classical AQA framework to the ASD dataset and explore the feasibility of AQA as a substitute for behavioural observation in ASD diagnosis.

3 Methodology

3.1 Network Design

Convolutional Neural Networks (CNN) or Recurrent Neural Networks (RNN) are frequently used to capture spatio-temporal information in skeletal sequences. To efficiently use image-based representation to capture the spatio-temporal information of a skeleton sequence presents a problem for CNN-based methods. We consider the benefits of a feature representation of skeletal sequence data in terms of its adaptability to perspective shifts and the postural characteristics that visualise the process of change in human movement. In order to combine the advantages of CNN in spatial structure information and RNN in temporal feature modelling. Therefore, we use a hybrid model of CNN and LSTM to obtain motion information. CNN is used for feature extraction to encode spatial information, while LSTM captures the temporal information of a sequence. The aim is to create a video-level description by combining sub-sequence skeleton characteristics to simulate temporal effects through the LSTM layer.

Considering the skeletal data structure, we chose 1D CNN as the backbone for feature extraction. And different size time windows were achieved using different kernel size settings.

As shown in Figure 2, for 3D skeletal sequence data with size $(F, N, 3)$, where F refers to the number of frames, and N is the number of skeleton joints. after reshaping, the data will be formed as $(F, N * 3)$. The 1D CNN will be convoluting the frame channel, and the filter size defines the size of the window for processing how many frames at a time, which refers to the clips at Figure 1.

3.2 Performance Metrics

In order to anticipate the quality "score," action quality assessment is presented as a regression problem. As a result, Spearman rank correlation is utilised to evaluate performance. It assesses how well the relationship between two variables can

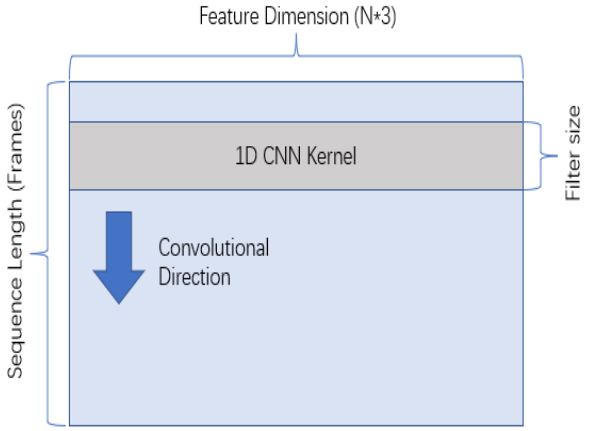

FIGURE 2. 1-Dimensional convolution illustration for skeleton sequence data

be described using a monotonic function. The Spearman's rank coefficient of correlation, or simply the Spearman's correlation coefficient, is a non-parametric measure of rank correlation. It reflects the correlation between the direction and intensity of the trend of two random variables.

$$p = 1 - \frac{6 \sum d_2^i}{n(n^2 - 1)} \tag{1}$$

where, p = Spearman's rank correlation coefficient; d_i = the difference between the two ranks of each observation; n = the number of observations. A higher p denotes a better rank correlation between the true and forecasted scores. This metric permits non-linear relationships, but it clearly stresses relative ordering rather than the true score value (i.e. lower scores for poor examples and higher scores for better quality examples)

4 Experiment and Results

4.1 DREAM Dataset

The DREAM dataset consists of a total of 306 hours of interaction data collected. The interaction set consisted of Robot Enhanced Therapy (RET) and standard human treatment (SHT) settings. And the two interaction configurations were designed to be as similar as possible, with the interaction partner constituting the primary difference. There were 61 participating children, and the clinical programme consisted of an initial assessment, eight interventions (3 imitation tasks, three joint attention tasks, and two turn-taking tasks) and a final assessment.

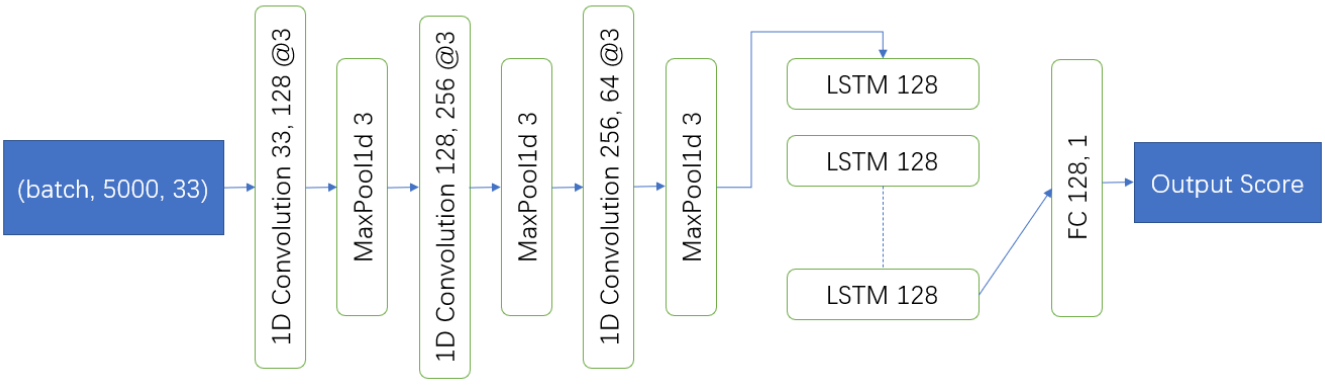

FIGURE 3. The Proposed network structure

The total length of each intervention ranged from 3 to 87 minutes, with a median duration of 32 minutes. A professional ADOS assessment was also given for the first and last assessments.

TABLE 1. The released data of DREAM dataset

Index	Release of the DREAM dataset
1.	Child ID (numerical index)
2.	Gender
3.	Age (in months)
4.	3D skeleton joints position for upper body
5.	3D head position and orientation
6.	3D eye gaze vectors
7.	Interaction condition (RET or SHT)
8.	Interaction task
9.	ADOS-G scores
10.	Date and duration of recording

The training label is obtained through the use of the ADOS-G scoring algorithms, which evaluate communication, social interaction, play, and stereotype scores. The ADOS-G provides a standardized method for observing current social-communicative behaviour, and has demonstrated strong inter-rater reliability, internal consistency, and test-retest reliability across items, domains, and categories for individuals with ASD and those without the ASD [14]. Those code domains for ADOS Module include the following assessment entries:

- A Domain communication: The overall extent of non-parody language; Frequency of vocalisation towards others; Intonation of vocal or spoken language; Immediate imitation of speech; Stereotypical/idiosyncratic use of single words or phrases; Uses other people's bodies to communicate; Pointing actions; Postural actions.

- B Domain social-interaction: Unusual eye contact; Reactive social smiling; Facial expressions towards others; Integration of eye gaze and other behaviours in the active expression of social intentions; Sharing fun during interaction; Response to name; Demanding; Giving; Demonstrating; Spontaneous and active production of reciprocal co-ordinated attention; Responses to reciprocal co-ordinated attention; Active expression of social intention.

- C Domain play: Functional play with objects; Imagination/creativity.

- D Domain Stereotyped Behaviours and Restricted Interests: Unusual sensory interest in play material/people; Hand and finger and other complex and specific movements of habit; Self-harming behaviour; Unusual repetitive interests or stereotyped behaviours;

Each scored domain is rated on a 4-point scale from 0 to 3, where a score of 2 or 3 indicates that the item has a definite abnormality, a score of 1 indicates a lesser abnormality, and 0 is no abnormality. Given that we aimed to use the behavioural data to build a regression model with the scoring criteria, after comparing the four domains in detail, we selected those four scores, communication, social interaction, play, and stereotypical behaviour, as the labels for our proposed AQA task.

4.2 Data Preprocessing

The behavioural data in the Dream dataset included 3-dimensional skeletal data for 10 joints Figure 5 and 3-dimensional gaze estimates. Considering that ADOS rating criteria are related to gaze behaviour. Therefore we combined the eye gaze data with the skeletal for training. Specifically, We performed a zero-fill operation on the gaze data's null data, then concatenate the gaze data and skeletal data together. Then, we linearly interpolated and scaled all sequences in the frame dimension so that all sequences were the same length.

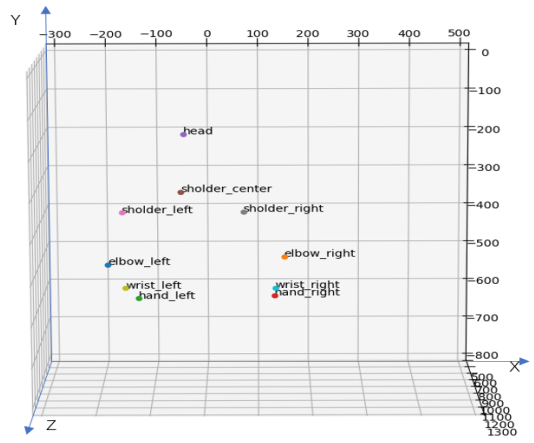

FIGURE 4. Skeleton data visualisation of DREAM dataset

As the DREAM dataset contains three different interaction tasks, each containing different difficulty levels, we only chose the data from Turn-Taking (TT) task with difficulty level 1 (Contains the most available data) to avoid excessive differences in the dataset to the greatest extent possible. A total of 149 interaction sequences were included, of which 70% were randomly selected for training and 30% for testing.

4.3 Implement Details

We implement our framework using PyTorch, the network is illustrated in Figure 3 The Adam optimiser with an initial learning rate of 0.0001 and weight decay of 0.0005, with ReduceLROnPlateau scheduler with a gamma of 0.1, patience of 5 epochs with minimum learning rate 5e-6 is employed to train the AQA model, the maximum epochs have set to 200.

TABLE 2. Spearman's rank correlation for ADOS scores

	Spearman's Rank Correlation			
Filter size	Com	Int	Play	Ste
3	0.262	0.535	0.251	0.230
6	0.223	0.344	0.387	0.126
12	0.266	0.349	0.283	0.261
18	0.262	0.353	0.392	0.133
30	0.338	0.297	0.415	0.376
60	0.321	0.313	0.348	0.326

4.4 Spearman's Rank Correlation for ADOS Scores

Performance comparison between 4 ADOS scoring domains is provided for 6 different sizes of filters on the DREAM dataset in Table. 2. The highest p value over those 4 domains occurs in the interaction domain ($p = 0.535$) with filter size = 3, and the second highest in the play domain with p value 0.415. According to the Spearman correlation coefficient, p values great than 0.6 are considered strongly correlated, between [0.4-0.6] are generally considered moderately correlated, and [0.2-0.4] are weakly correlated. Therefore, our AQA model results are moderately correlated with the ground truth labels for interaction and play domains. However, compared with interaction and play, the p value for communication and stereotype domains only obtained a weak correlation based on the result.

Although we do not observe strong correlations with p-values greater than 0.6 in the table2, the DREAM dataset is not specifically designed for the AQA tasks. Even though specialist clinicians assess the ADOS scores, the assessment subcategories are not strictly correlated with the behavioural data, and the dataset does not provide detailed sub-category assessment item scores for each of the four domains.

5 Summary

In ASD diagnosis and early detection, behaviour observation by clinicians has become its bottleneck. This paper attempts to use machine learning-based behaviour quantified methods to provide biomarkers to speed up the behaviour observation process for ASD diagnosis and early detection. We introduced the AQA approach to the field of ASD research and used the existing ASD behaviour dataset to initially demonstrate the feasibility of the AQA model for replacing manual observation scores. Our finding provides a new direction for future computer-assisted observation of ASD behaviour. Despite its potential, this work is still in its preliminary state. The proposed approaches for quantifying behaviour will be examined

in future studies to assess their potential for ASD diagnosis and clinical rehabilitation.

References

[1] S. Wang, D. Yang, P. Zhai, Q. Yu, T. Suo, Z. Sun, K. Li, and L. Zhang, "A survey of video-based action quality assessment," in *2021 International Conference on Networking Systems of AI (INSAI)*. IEEE, 2021, pp. 1–9.

[2] P. Parmar and B. T. Morris, "What and how well you performed? a multitask learning approach to action quality assessment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[3] X. Yu, Y. Rao, W. Zhao, J. Lu, and J. Zhou, "Group-aware contrastive regression for action quality assessment," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 7919–7928.

[4] P. Parmar and B. Tran Morris, "Learning to score olympic events," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2017, pp. 20–28.

[5] E. Billing, T. Belpaeme, H. Cai, H.-L. Cao, A. Ciocan, C. Costescu, D. David, R. Homewood, D. Hernandez Garcia, P. Gómez Esteban *et al.*, "The dream dataset: Supporting a data-driven study of autism spectrum disorder and robot enhanced therapy," *PloS one*, vol. 15, no. 8, pp. 1–15, 2020.

[6] K. Gotham, A. Pickles, and C. Lord, "Standardizing ados scores for a measure of severity in autism spectrum disorders," *Journal of autism and developmental disorders*, vol. 39, no. 5, pp. 693–705, 2009.

[7] H. Pirsiavash, C. Vondrick, and A. Torralba, "Assessing the quality of actions," in *European conference on computer vision*. Springer, 2014, pp. 556–571.

[8] P. Parmar and B. Morris, "Action quality assessment across multiple actions," in *2019 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 2019, pp. 1468–1476.

[9] C. Xu, Y. Fu, B. Zhang, Z. Chen, Y.-G. Jiang, and X. Xue, "Learning to score figure skating sport videos," *IEEE transactions on circuits and systems for video technology*, vol. 30, no. 12, pp. 4578–4590, 2019.

[10] Y. Gao, S. S. Vedula, C. E. Reiley, N. Ahmidi, B. Varadarajan, H. C. Lin, L. Tao, L. Zappella, B. Béjar, D. D. Yuh, C. C. G. Chen, R. Vidal, S. Khudanpur, and G. Hager, "Jhu-isi gesture and skill assessment working set (jigsaws) : A surgical activity dataset for human motion modeling," in *MICCAI workshop: M2cai*, 2014, pp. 1–10.

[11] 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*, 2018, pp. 6057–6066.

[12] 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/CVF conference on computer vision and pattern recognition*, 2019, pp. 7862–7871.

[13] Z. Li, Y. Huang, M. Cai, and Y. Sato, "Manipulation-skill assessment from videos with spatial attention network," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019, pp. 1–11.

[14] C. Lord, M. Rutter, P. C. DiLavore *et al.*, "Autism diagnostic observation schedule–generic," *Dissertation Abstracts International Section A: Humanities and Social Sciences*, 1999.