2021 IEEE International Conference on Image Processing (ICIP) | 978-1-6654-4115-5/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICIP42928.2021.9506257

ACTION QUALITY ASSESSMENT WITH IGNORING SCENE CONTEXT

Takasuke Nagai, Shoichiro Takeda, Masaaki Matsumura, Shinya Shimizu, Susumu Yamamoto

NTT Media Intelligence Laboratories, NTT Corporation, Japan

ABSTRACT

We propose an action quality assessment (AQA) method that can specifically assess target action quality with ignoring scene context, which is a feature unrelated to the target action. Existing AQA methods have tried to extract spatiotemporal features related to the target action by applying 3D convolution to the video. However, since their models are not explicitly designed to extract the features of the target action, they mis-extract scene context and thus cannot assess the target action quality correctly. To overcome this problem, we impose two losses to an existing AQA model: scene adversarial loss and our newly proposed human-masked regression loss. The scene adversarial loss encourages the model to ignore scene context by adversarial training. Our human-masked regression loss does so by making the correlation between score outputs by an AQA model and human referees undefinable when the target action is not visible. These two losses lead the model to specifically assess the target action quality with ignoring scene context. We evaluated our method on a diving dataset commonly used for AQA and found that it outperformed current state-of-the-art methods. This result shows that our method is effective in ignoring scene context while assessing the target action quality.

Index Terms—Action quality assessment, Scene context, Deep learning, Spatiotemporal feature, Regression problem

1. INTRODUCTION

Action quality assessment (AQA) is the task of automatically assessing target action quality in an input video. AQA has many applications, such as assisting sports judgements [10], monitoring surgical skills [4, 26], checking daily behavior [4, 5], and enhancing healthcare [6, 8]. In particular, its use in sports has attracted attention because it improves the fairness of judgements by ensuring impartiality [1, 27].

For AQA, several methods have proposed the utilization of human pose features such as joint positions and/or joint motions [13, 17, 22]. However, such methods are difficult to implement in real situations because pose estimation accuracy is limited for complex actions such as those in diving and gymnastics, where joints are often occluded. In fact, state-of-the-art pose estimation methods cannot deal with the joint occlusions well [3, 19, 25], and thus their incorrect estimation results degrade AQA performance [17]. On the other hand, several methods utilize spatiotemporal features extracted from the video by using a deep neutral network (DNN) model based on 3D convolution [11, 12, 14–16, 20, 23, 24]. These methods are promising because they do not rely on the pose estimation accuracy and they can achieve comparable or superior accuracy to the pose feature-based methods. However, since their models are not explicitly designed to extract the features of the target action, they sometimes mis-extract scene context. We define scene context here as background features unrelated to the target action, such as the color/shape/motion features of background objects (including the audience). For example, Fig. 1 shows a correlation between scores output by an AQA model and human referees. We used the state-of-the-art C3D-AVG model (surrounded by black) here, proposed by Parmar *et al*. [16]. This model consists of a spatiotemporal feature extractor called the C3D model [21] and an averaging (AVG) operator for aggregating the spatiotemporal features. The C3D-AVG model had a high correlation for normal videos, but it also provided highly correlated scores for human masked videos even though the target action was completely invisible. This result shows that the C3D-AVG model unfortunately performed AQA with mis-extracted scene context rather than the features of the target action. We therefore consider that this model is inappropriate for AQA due to its unnatural assessment of the human-masked videos.

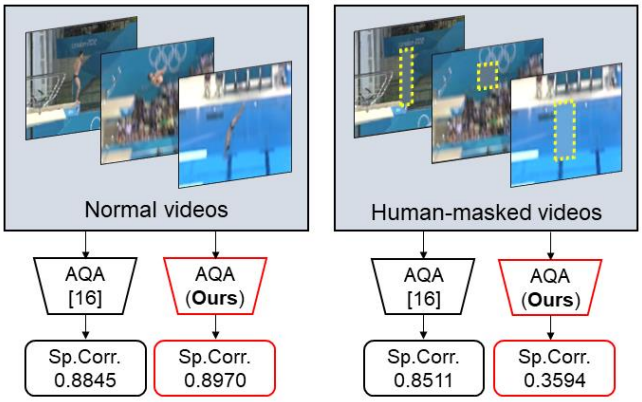

Fig. 1. Results of input video assessments. (Left) Inputting normal videos. (Right) Inputting human-masked videos, in which the target action is completely masked. Although the target action was not visible in the human-masked videos, the existing C3D-AVG model [16] had surprisingly high correlations with the ground-truth human referee's score. However, our proposed model could specifically assess the target action quality with ignoring scene context. The accuracy was evaluated using the Spearman's rank correlation (Sp. Corr.).

For improving AQA performance, we propose an AQA method that can specifically assess the target action quality with ignoring scene context. Our method imposes two losses to the state-of-the-art C3D-AVG model: scene adversarial loss [2] and our original human-masked regression loss. The scene adversarial loss is used to encourage the model to ignore scene context in the input video by adversarial training, and our human-masked regression loss does so by making the AQA score correlation undefinable when the target action is not visible. Our human-masked regression loss is introduced with the same concept as the human mask confusion loss [2]; we have reformulated to fit regression problems such as AQA.

Based on the idea of maximizing cross-entropy in the human mask confusion loss to make the action label unpredictable, our loss regresses all estimated scores toward the same fixed score of 0 for making the correlation undefinable when the target action is not visible. Note that this regression toward 0 assumes that the human referee will assess no target action as a 0 quality score.

Our contributions can be summarized as follows. 1) We propose an AQA method that specifically assesses the target action quality with ignoring scene context by newly imposing two losses: scene adversarial loss and our original human-masked regression loss. 2) We developed the human-masked regression loss for the regression problem inspired by the maximizing cross-entropy idea of the action classification problem. 3) We demonstrate that our method outperforms existing methods on the MTL-AQA dataset [16] and that ignoring scene context is effective for assessing the target action quality.

2. EXISTING METHOD

Parmar *et al*. proposed an AQA method with the C3D-AVG model [16], which utilizes the spatiotemporal features in the video to output the target action quality score. In this method, a normal video with 96 frames is divided into six 16-frames clips. Then, for extracting clip-level spatiotemporal features, each clip is input to a weight-shared C3D [21] without the last two fully connected (FC) layers. The clip-level spatiotemporal features are averaged to obtain video-level spatiotemporal features. The video-level spatiotemporal features are finally passed into an action-quality-score estimator to output the target action quality score. The overview of this method is shown in Fig. 2(a).

To estimate the target action quality score from the spatiotemporal features, Parmar *et al*. [16] considered AQA as a regression problem and minimized the score regression loss as

$$L_{SR}(\boldsymbol{x_n}, y_n, \boldsymbol{\theta}, \boldsymbol{\Phi}) = (\hat{y}_n - y_n)^2 + |\hat{y}_n - y_n|, \quad \hat{y}_n = A(F(\boldsymbol{x_n};\, \boldsymbol{\theta});\, \boldsymbol{\Phi}), \tag{1}$$

where $\boldsymbol{x_n}$ is an n-th normal video, y_n is an n-th ground-truth target action quality score provided by human referees, $\hat{y}_n$ is an n-th estimated target action quality score by the C3D-AVG model, $F(\cdot;\boldsymbol{\theta})$ is the feature extractor in the C3D-AVG model parameterized by $\boldsymbol{\theta}$, and $A(\cdot;\boldsymbol{\phi})$ is the action-quality-score estimator in the C3D-AVG model parameterized by $\boldsymbol{\phi}$. By minimizing Eq. (1), the C3D-AVG model learns to estimate the target action quality score with the spatiotemporal features in the video. However, since this model is not explicitly designed to extract the features of the target action, it mis-extracts scene context (as shown in Fig. 1). In this paper, we propose a method that can specifically assess the target action quality with ignoring scene context.

3. PROPOSED METHOD

In addition to Eq. (1), our method imposes two losses to the existing C3D-AVG model: scene adversarial loss [2] and our original human-masked regression loss. The details of the losses and optimization are described below. The overview of our method is shown in Fig. 2(b).

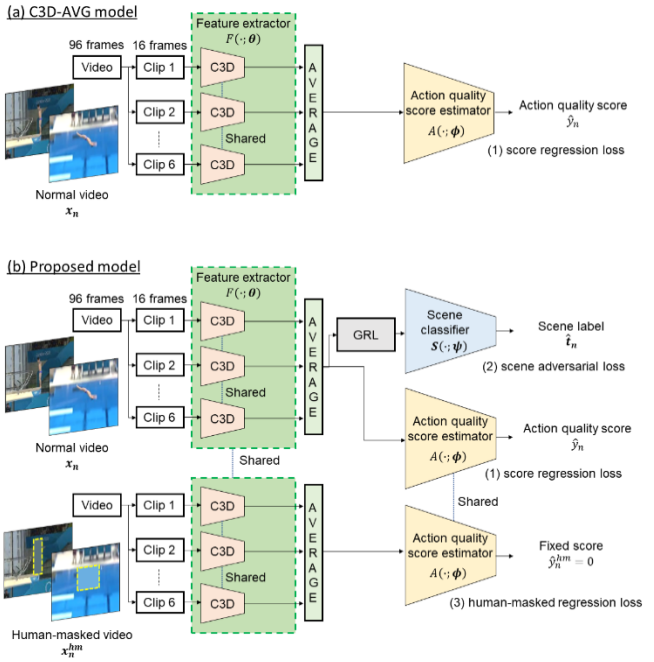

Fig. 2. Overview of (a) conventional C3D-AVG model [16] and (b) proposed model. There are two types of input video in our method. (First row in proposed model) With normal video $\boldsymbol{x_n}$, the scene adversarial loss is imposed in addition to the score regression loss. (Second row in proposed model) With human-masked video $\boldsymbol{x_n^{hm}}$, the human-masked regression loss is imposed. In this figure, each mask is surrounded by a yellow dot rectangle only for visualization. This figure was created with reference to [2], [16].

3.1. Scene Adversarial Loss

In addition to Eq. (1), we impose the scene adversarial loss [2] to the C3D-AVG model for ignoring scene context in AQA. The scene adversarial loss discourages the model from extracting the spatiotemporal features related to scene context by learning the scene label in an adversarial manner. This loss consists of a cross-entropy between the estimated scene label, which is the output of the scene classifier with the video-level spatiotemporal features input, and the ground-truth one. We define the scene adversarial loss as

$$L_{SA}(\boldsymbol{x_n}, \boldsymbol{t_n}, \boldsymbol{\theta}, \boldsymbol{\psi}) = \frac{1}{K}\sum_{k=1}^{K} -t_{n,k}\, log\, \hat{t}_{n,k},$$

$$\hat{\boldsymbol{t}}_{\boldsymbol{n}} = \left[\hat{t}_{n,1}, \ldots, \hat{t}_{n,K}\right]^{\top} = \boldsymbol{S}\left(R_{\lambda}\left(F(\boldsymbol{x}_n;\boldsymbol{\theta})\right);\boldsymbol{\psi}\right),$$

$$\boldsymbol{t_n} = \left[t_{n,1}, \ldots, t_{n,K}\right]^{\top}, \tag{2}$$

$$R_{\lambda}\left(F(\boldsymbol{x}_n;\boldsymbol{\theta})\right) = F(\boldsymbol{x}_n;\boldsymbol{\theta}),$$

$$\frac{\partial R_{\lambda}\left(F(\boldsymbol{x}_n;\boldsymbol{\theta})\right)}{\partial \boldsymbol{\theta}} = -\lambda \frac{\partial F(\boldsymbol{x}_n;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}},$$

where $\boldsymbol{t_n}$ is an n-th ground-truth scene label with K classes, $\hat{\boldsymbol{t}}_{\boldsymbol{n}}$ is an n-th estimated scene label by our proposed model, $\boldsymbol{S}(\cdot;\boldsymbol{\psi})$ is the scene classifier in the model parameterized by $\boldsymbol{\psi}$, $R_{\lambda}(\cdot)$ is the gradient reversal layer (GRL), which inverts the sign of the gradient

during back-propagation, and λ is a hyper-parameter for the gradient of GRL. Note that the ground-truth scene label represents the input video scene venue, such as the London Aquatics Center or Nanjing Olympic Sports Center.

By minimizing Eq. (2), $\boldsymbol{S}(\cdot;\boldsymbol{\psi})$ learns to classify the scene label but $F(\cdot;\boldsymbol{\theta})$ learns adversarially so as not to extract the features of the scene label, which is related to scene context, by GRL. Therefore, $F(\cdot;\boldsymbol{\theta})$ is encouraged to ignore scene context.

3.2. Human-masked Regression Loss

In addition to the scene adversarial loss, we propose a new loss called human-masked regression loss for ignoring scene context in AQA. The important point of this regression loss is calculated by utilizing the output scores when inputting human-masked videos to the model. The human-masked regression loss encourages the model to make the correlation between scores output by the model and human referees undefinable when the target action is not visible. Thus, we define the human-masked regression loss for regressing the estimated scores toward the fixed scores as

$$L_{HMreg}\left(\boldsymbol{x_n^{hm}}, y_n^{hm}, \boldsymbol{\theta}, \boldsymbol{\Phi}\right) = (\hat{y}_n^{hm} - y_n^{hm})^2 + \left|\hat{y}_n^{hm} - y_n^{hm}\right|, \qquad \hat{y}_n^{hm} = A(F(\hat{\boldsymbol{x}}_{\boldsymbol{n}};\, \boldsymbol{\theta});\, \boldsymbol{\Phi}), \tag{3}$$

where $\boldsymbol{x_n^{hm}}$ is an n-th human-masked video, y_n^{hm} is an n-th fixed score, and $\hat{y}_n^{hm}$ is an n-th estimated score by our model.

This new loss is introduced with the same concept as the human mask confusion loss proposed by Choi *et al*. [2]; we reformulate the loss to fit the regression problem in AQA. The human mask confusion loss [2] aims to make the action label unpredictable when the target action is not visible by maximizing the cross-entropy between the predicted distributions of the action label in the action classification problem. As a result, the model loses action classification ability only for scene context and is encouraged to extract the features of only the target action. On the other hand, in the regression problem, we assume that the idea of making the correlation undefinable when the target action is not visible is equivalent to that of maximizing cross-entropy, as both ideas make the target action quality/label unpredictable. On the basis of this idea, we regress all estimated scores toward the same fixed score to make the correlation undefinable and thus ignore scene context. Note that when all estimated scores are the same, the correlation is undefinable because calculating it is impossible. In this paper, we set the fixed score y_n^{hm} to 0 by assuming that when the target action is not visible in the human-masked video, the referee cannot assess action quality and give a 0 quality score.

By minimizing Eq. (3), the model learns to make the correlation undefinable when the target action is not visible. As a result, the model is encouraged to ignore scene context for assessing the target action quality.

3.3. Optimization

To specifically assess the target action quality with ignoring scene context, we define the optimization problem using Eqs. (1)–(3) as

$$\min_{\boldsymbol{\theta},\boldsymbol{\phi},\boldsymbol{\psi}} \sum_{n=1}^{N} L_{SR}(\boldsymbol{x}_n, y_n, \boldsymbol{\theta}, \boldsymbol{\phi}) + \alpha L_{SA}(\boldsymbol{x_n}, \boldsymbol{t_n}, \boldsymbol{\theta}, \boldsymbol{\psi}) + \beta L_{HMreg}\left(\boldsymbol{x_n^{hm}}, y_n^{hm}, \boldsymbol{\theta}, \boldsymbol{\phi}\right), \tag{4}$$

where α and β are hyper-parameters that tune the weight of the loss L_{SA} and L_{HMreg}, respectively. When inputting normal videos, we minimize only L_{SR} and L_{SA}, and this optimization leads to adversarial training, where $F(\cdot;\boldsymbol{\theta})$ learns to extract the features of the target action that can be used to estimate the target action quality score but not to classify the scene label. In addition, when inputting human-masked videos, we minimize only L_{HMreg}, and this optimization leads to making the AQA score correlation undefinable when the target action is not visible. As a result, by minimizing Eq. (4), the model can specifically assess the target action quality with ignoring scene context.

4. EXPERIMENT

First, we describe the diving dataset [16] used in this experiment and the implementation details. Then, we tested the proposed AQA method could focus on the target action quality with ignoring scene context in the input video by comparing its accuracy with the state-of-the-art methods. Finally, we evaluated the effectiveness of the proposed method with various fixed score settings.

4.1. Dataset

We used the MTL-AQA dataset [16], which is the largest publicly available diving dataset. It includes 1412 video clips that are each annotated with a score, action class, and captions as ground-truth labels. The dataset consists of five scenes in different sports venues, including the London Aquatics Center and Nanjing Olympic Sports Center. Therefore, we set the number of scene labels *K* in Eq. (2) to five. For the experiment in Sections 4.3 and 4.4, we randomly chose 80% of the 1412 videos as training data, 10% as validation data, and 10% as test data. Note that only the results of test data are described.

4.2. Implementation Details

We implemented all DNN models using PyTorch (version 1.2.0). C3D-AVG [16], on which our model is based, was used as a baseline. As a feature extractor, all models described in the experiment used the C3D model [21] with weights that were pre-trained on the UCF101 dataset [18]. The hyper-parameters in Eqs. (2) and (4) were $\alpha = 0.01$, $\beta = 0.01$, and $\lambda = 0.01$. All models were learned using the Adam optimizer for 100 epochs with the initial learning rate of 1e-6. The resolution of the input videos was 171 × 128 pixels, and we applied a center crop of 112 × 112 pixels. During the learning process, a random horizontal flipping of the input video was performed. The batch size was set to three. For creating the human-masked videos, the masks were initially detected by using Mask R-CNN [9] and then manually adjusted to cover just the target action of a competitor. The mask of the human-masked video was colored using the average pixel value of each frame in the same manner as Choi *et al*. [2]. The accuracy was evaluated using the Spearman's rank correlation (Sp. Corr.) in accordance with the existing method [11–17, 20, 22–24].

Table 1. Results of input normal videos and human-masked videos.

Model	Sp. Corr.		Diff
	Normal↑	Masked↓	
Li *et al*. [11]	0.8755	0.8230	0.0525
C3D-SVR [14]	0.6977	0.6326	0.0651
C3D-LSTM [15]	0.8331	0.7634	0.0697
C3D-AVG [16] (Baseline)	0.8845	0.8511	0.0334
C3D-AVG-SA	0.8875	0.8525	0.0350
C3D-AVG-HMreg	0.8964	0.3607	0.5357
Ours (C3D-AVG-SA&HMreg) w/ predicted mask	0.8949	**0.3398**	**0.5551**
Ours (C3D-AVG-SA&HMreg) w/ manually adjusted mask	**0.8970**	0.3594	0.5376

4.3. Effectiveness of Proposed Method

AQA performance on normal videos. In this experiment, we tested whether the proposed AQA method could assess the target action quality correctly. In this experiment, the correlation should be high because the target action is present. The second column of Table 1 shows the results of our method in comparison with the C3D-based state-of-the-art methods: Li *et al*. [11], C3D-SVR [14], C3D-LSTM [15], and C3D-AVG [16] (Baseline). Note that, for fair comparison, we prepared the models learned with only the scene adversarial loss, Eq. (2), or the human-masked regression loss, Eq. (3), as C3D-AVG-SA or C3D-AVG-HMreg. The table shows that both C3D-AVG-SA and C3D-AVG-HMreg improved the correlation relative to the Baseline. Furthermore, we prepared our models (C3D-AVG-SA&HMreg) learned with the predicted masks by Mask R-CNN without manual adjustment, or the manually adjusted masks to cover just the target action of a competitor. Both of our models outperformed all state-of-the-art models even with the predicted masks, and the model with manually adjusted mask showed improvements relative to C3D-AVG-SA and C3D-AVG-HMreg. These results demonstrate that our model can correctly assess the target action quality by imposing two losses to the C3D-AVG model.

AQA performance on human-masked videos. In this experiment, we tested whether the proposed AQA method could ignore scene context in the video. The third column of Table 1 shows the AQA performance for the human-masked videos, in which the target action is completely masked, and the fourth column of Table 4 shows the correlation difference between normal video and human-masked video. In this experiment, the correlation should be low because the target action is absent, and the correlation difference should be large because the correlation of normal video should be high and that of human-masked video should be low. Curiously, compared to the performance for the normal video, there was only a slight performance difference with the state-of-the-art models even though the target action was not visible in the human-masked videos; these models mis-extracted scene context during optimization, so they might be inappropriate for AQA due to unfair judgement caused by scene context. In contrast, C3D-AVG-HMreg and both of our models showed significant drops compared with the performance for normal videos even with the predicted masks. This means that our introducing loss, Eq. (3), is effective for ignoring scene context. Furthermore, the loss of Eq. (2) also encourages the model slightly but surely to ignore scene context because this loss succeeds in widening the correlation difference from Baseline to C3D-AVG-SA and from C3D-AVG-HMreg to Ours. These results demonstrate that our model can specifically assess the target action quality with ignoring scene context.

Table 2. Results of learning various fixed scores when inputting normal videos and human-masked videos.

Fixed score	Sp. Corr.	
	Normal↑	Masked↓
100	0.8887	0.7178
90	0.8930	0.7697
80	0.8955	0.8182
70	0.8976	0.8426
67 (Average)	0.8988	0.8434
60	0.8993	0.8429
50	0.8991	0.8022
40	0.8985	0.7278
30	**0.8999**	0.6322
20	0.8982	0.5358
10	0.8980	0.4453
0 (Ours)	0.8970	**0.3594**

4.4. Effectiveness of Proposed Fixed Score Setting

In this experiment, we tested the effect of the fixed score on AQA performance. Table 2 shows the results with various fixed scores during the model learning when inputting normal videos and human-masked videos. These fixed scores were selected from 0 to 100, which is within the score range of the dataset (0 to 104.5). "Average" means the average score of the dataset. For the normal videos, high correlation was achieved when the fixed value was 30, 50, and 60, and the correlation decreased as the fixed score increased or decreased. On the other hand, for the human-masked videos, a significant drop was evident when the fixed score was 0. The fixed score of 70 provided the smallest performance difference between the normal and human-masked video inputs, and the difference increased as the fixed score increased or decreased. Considering the AQA requirement of assessing just the target action, low correlation is better when inputting human-masked video, and the fixed score of 0 is logical because a referee cannot assess action quality when the target action is not visible. Therefore, considering a balance between the AQA performance and AQA requirement, we conclude that the fixed score of 0 is the most appropriate for AQA.

5. CONCLUSION

We proposed an AQA method that can specifically assess the target action quality with ignoring scene context by newly imposing two losses: the scene adversarial loss and our original human-masked regression loss. The former encourages the model to ignore scene context by adversarial training, and the latter does so by making the AQA score correlation undefinable when the target action is not visible. Our human-masked regression loss is minimized to regress all the estimated scores toward the same fixed score of 0. Our concept is based on the idea that maximizing cross-entropy between the predicted action label distributions stands for making a model lose prediction ability on an action classification problem. The experimental results show that our method outperformed the existing methods, and that ignoring scene context is effective to specifically assess the target action quality. In future work, to verify the generality of the proposed method, we should investigate how the correctness and shape of the mask affect the AQA performance.

6. REFERENCES

[1]. M.J. Borzilleri. Olympic Figure Skating Controversy: Judging System is Most to Blame for Uproar, 2014.

[2]. J. Choi, C. Gao, J.C.E. Messou, and J.B. Huang. Why Can't I Dance in the Mall? Learning to Mitigate Scene Bias in Action Recognition. In NeurIPS, 2019.

[3]. X. Chu, W. Yang, W. Ouyang, C. Ma, A.L. Yuille, and X. Wang. Multi-context attention for human pose estimation. In CVPR, 2017.

[4]. H. Doughty, D. Damen, and W.M. Cuevas. Who's Better, Who's Best: Skill Determination in Video Using Deep Ranking. In CVPR, 2018.

[5]. H. Doughty, W.M. Cuevas, and D. Damen. The Pros and Cons: Rank-aware Temporal Attention for Skill Determination in Long Videos. In CVPR, 2019.

[6]. T. Elgamal and K. Nahrstedt. Multivariate Multicamera Summarization of Rehabilitation Sessions in Home Environment. In ACM Multimedia, 2017.

[7]. Y. Ganin and V. Lempitsky. Unsupervised domain adaptation by backpropagation. In ICML, 2015.

[8]. T. Hakim and I. Shimshoni. A-MAL: Automatic Motion Assessment Learning from Properly Performed Motions in 3D Skeleton Videos. In CVPM, 2019.

[9]. K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask R-CNN. In ICCV, 2017.

[10]. Q. Lei, J.X. Du, H.B. Zhang, and S. Ye. A Survey of Vision-Based Human Action Evaluation Methods. Sensors, 19(19), 2019.

[11]. Y. Li, X. Chai, and X. Chen. End-To-End Learning for Action Quality Assessment. In Pacific Rim Conference on Multimedia, 2018.

[12]. Y. Li, X. Chai, and X. Chen. ScoringNet: Learning Key Fragment for Action Quality Assessment with Ranking Loss in Skilled Sports. In ACCV, 2018.

[13]. J.H. Pan, J. Gao, and W.S. Zheng. Action Assessment by Joint Relation Graphs. In ICCV, 2019

[14]. P. Parmar and B.T. Morris. Learning to Score Olympic Events, In CVPR Workshop, 2017

[15]. P. Parmar and B.T. Morris. Action Quality Assessment Across Multiple Actions. In WACV, 2019.

[16]. P. Parmar and B.T. Morris. What and How Well You Performed? A Multitask Learning Approach to Action Quality Assessment. In CVPR, 2019.

[17]. H. Pirsiavash, C. Vondrick, and A. Torralba. Assessing the Quality of Actions. In ECCV, 2014.

[18]. K. Soomro, A.R. Zamir, and M. Shah. Ucf101: A Dataset of 101 Human Actions Classes from Videos in the Wild. CRCV-TR-12-01, 2012.

[19]. K. Sun, B. Xiao, D. Liu, and J. Wang. Deep high-resolution representation learning for human pose estimation. In CVPR, 2019.

[20]. Y. Tang, Z. Ni, J. Zhou, D. Zhang, J. Lu, Y. Wu, and J. Zhou. Uncertainty-aware Score Distribution Learning for Action Quality Assessment. In CVPR, 2020

[21]. D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning Spatiotemporal Features with 3D Convolutional Networks. In ICCV, 2015.

[22]. V. Venkataraman, I. Vlachos, and P. Turaga. Dynamical Regularity for Action Analysis. In BMVC, 2015.

[23]. X. Xiang, Y. Tian, A. Reiter, G.D. Hager, and T.D. Tran. S3D: Stacking Segmental P3D for Action Quality Assessment. In ICIP, 2018.

[24]. 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, 2019.

[25]. W. Yang, S. Li, W. Ouyang, H. Li, and X. Wang. Learning feature pyramids for human pose estimation. In ICCV, 2017.

[26]. A. Zia, Y. Sharma, V. Bettadapura, E.L. Sarin, T. Ploetz, M.A. Clements, and I. Essa. Automated Video-Based Assessment of Surgical Skills for Training and Evaluation in Medical Schools. International journal of computer assisted radiology and surgery, 11(9):1623–1636, 2016.

[27]. Wikipedia contributors. List of Olympic Games Scandals and Controversies — Wikipedia, the free encyclopedia, 2021. [Online; accessed 13-January-2021].