Uncertainty-aware Score Distribution Learning for Action Quality Assessment

Yansong Tang\textsuperscript{1,2,3,\ast}, Zanlin Ni\textsuperscript{1,\ast}, Jiahuan Zhou\textsuperscript{5}, Danyang Zhang\textsuperscript{1}, Jiwen Lu\textsuperscript{1,2,3}, Ying Wu\textsuperscript{5}, Jie Zhou\textsuperscript{1,2,3,4}

\textsuperscript{1} Department of Automation, Tsinghua University, China
\textsuperscript{2} State Key Lab of Intelligent Technologies and Systems, China
\textsuperscript{3} Beijing National Research Center for Information Science and Technology, China
\textsuperscript{4} Tsinghua Shenzhen International Graduate School, Tsinghua University, China
\textsuperscript{5} Electrical and Computer Engineering Department, Northwestern University

\{tys15,nzl17,zhang-dy16\}@mails.tsinghua.edu.cn; jiahuanzhou2013@u.northwestern.edu; \{lujiwen,jzhou\}@tsinghua.edu.cn; yingwu@eecs.northwestern.edu

Abstract

Assessing action quality from videos has attracted growing attention in recent years. Most existing approaches usually tackle this problem based on regression algorithms, which ignore the intrinsic ambiguity in the score labels caused by multiple judges or their subjective appraisals. To address this issue, we propose an uncertainty-aware score distribution learning (USDL) approach for action quality assessment (AQA). Specifically, we regard an action as an instance associated with a score distribution, which describes the probability of different evaluated scores. Moreover, under the circumstance where fine-grained score labels are available (e.g., difficulty degree of an action or multiple scores from different judges), we further devise a multi-path uncertainty-aware score distributions learning (MUSDL) method to explore the disentangled components of a score. We conduct experiments on three AQA datasets containing various Olympic actions and surgical activities, where our approaches set new state-of-the-arts under the Spearman’s Rank Correlation\textsuperscript{1}.

1. Introduction

Action quality assessment (AQA), aiming to evaluate how well a specific action is performed, has become an emerging and attractive research topic in computer vision community because of its potential value for various real-world applications such as sport video analysis \cite{11,18,19,22}, health care \cite{5,40,43} and many others \cite{3,4,42,44}. Compared with the conventional action recognition problem \cite{2,6,23,26,29,31,32,34,35,36} focusing on correctly classifying the action sequences from different categories, AQA is a more challenging task as it requires to deal with the videos from the same category with poor intra-class discriminant.

Over the past few years, there have been numbers of methods proposed for AQA \cite{12,13,17,18,21,22}. However, most of them simply regard AQA as a regression problem in order to directly predict the action scores. And unfortunately, their performances are indeed limited. The root of such a limitation is that this kind of treatment disregards for the underlying ambiguity of the action score labels, which is one of the key issues for AQA. Such ambiguity is caused by how the action labels are generated...
in practice. As an example shown in Figure 1 for the diving game, when the athlete finishes his action with the difficulty degree of 3.8, seven judges give their scores as \{9.0, 8.5, 9.0, 8.0, 9.0, 8.5, 9.0\}. After eliminating the top-two and bottom-two scores, the final score is calculated as: 

\[ s_{final} = (9.0 + 9.0 + 8.5) \times 3.8 = 100.70. \]

This suggests the inherent uncertainty of the final score caused by different judges. Moreover, the subjective appraisal of each individual judge might also bring uncertainty into the final score. Besides the diving game, these phenomena exist in many other sports like gymnastic vaults, figure skiing, etc. The complicated score uncertainty makes an accurate AQA pretty difficult. Hence, it is desirable to design a robust model to deal with the uncertainty for AQA.

To address this, we propose an uncertainty-aware score distribution learning (USDL) method which utilizes a distribution of different scores as the supervisory signal rather than a single score. The adopted score distribution can depict the probability of the AQA score better so that the aforementioned uncertainty issue can be well handled. As illustrated by the top-left of Figure 1, we generate the ground-truth score distribution based on the widely used Gaussian function, of which the mean is set to be the score label. Meanwhile, an action video is fed into a 3D ConvNets to produce its predicted score distribution. Then we optimize the kullback-leibler divergence between the ground-truth score distribution and the predicted one. Recall Figure 1 once the fine-grained score labels are available (e.g., difficulty degree of an action or multiple scores from different judges), we further design a multi-path uncertainty-aware score distributions learning (MUSDL) method to fully explore these disentangled components of the final score. Strictly abided by the rule of the game, we fuse the multiple predicted scores to obtain the final score during inference. Through this objective process, we are able to obtain more accurate results. To our best knowledge, this is the original effort to leverage the finer-level score annotation for AQA problem. In order to verify the effectiveness of our approach, we conduct extensive experiments on the AQA-7 [17], MTL-AQA [19] and JIGSAWS [8] datasets. Experimental results show the superiority of our methods compared with the state-of-the-arts, and demonstrate the advantage of utilizing fine-grained labels.

2. Related Work

Action Quality Assessment: In the past years, there have been a variety of works dedicated to different AQA tasks, such as health care [40], instructional video analysis [54], sport video analysis [10, 18, 19, 22] and many others [12, 15]. For example, Pirsiavash et al. [22] explored this task firstly based on hand-crafted features on several specific actions, they took the first step to apply the learning approach on the underlying task and trained a linear SVR model to regress the score of the videos. Parmar et al. [18] proposed C3D-SVR and C3D-LSTM to predict the score of the Olympic events. Additionally, incremental-label training method was introduced to train the LSTM model based on the hypothesis that the final score is an aggregation of the sequential sub-action scores. [17] and [19] aimed to improve the scoring performance and the generalization ability of the model simultaneously by exploring all-action models and multi-task learning respectively. Meanwhile, both of them released a new AQA dataset with a larger scale. Xu et al. [37] designed two new LSTM-based models to learn the multi-scale information of the video. A little different from the works focusing on extracting the whole-scene features, Pan et al. [16] presented a graph-based model to sufficiently exploit the athletes’ pose information. Their approach well balanced the role of the movement of body parts and the coordination among different joints. Different from the aforementioned regression-based AQA methods, our proposed USDL method aims to predict the score distribution for the input action video instead of a single score number so that the severe score uncertainty which largely limits AQA performance can be well handled.

Label Distribution Learning: Label distribution learning (LDL) is a general learning paradigm, which describes an instance with distribution rather than the original single label or multiple labels. As a pioneering work, Geng et al. [11] proposed an LDL framework for facial age estimation, which assigned an age distribution to each face image, and learned from such distribution with two algorithms called IIS-LLD (Improved Iterative Scaling-Learning from Label Distributions) and CPNN (Conditional Probability Neural Network). Motivated by the success of [11] which was based on the hand-crafted features, the following ameliorated works have been proposed for LDL by leveraging the power of deep learning models such as deep convolutional neural network [7] or LDL forest [25]. In recent years, LDL has also shown its effectiveness for various computer vision tasks, including head pose estimation [10], beauty sensing [24], facial landmark detection [27] and many others [36, 39, 41]. For video analysis, Geng et al. [9] proposed a soft grammar parsing method for video parsing, where the video segments are described by the degrees of different sub-action classes. Ling et al. [15] utilized a mixture of Gaussian distribution to model the gradual change of crowd numbers in different video frames for indoor crowd counting. For our proposed method, by transferring the given single score label to a Gaussian-like score distribution for learning, we are able to directly estimate the score distribution of an action video that provides more accurate AQA results than only predicting a single score.

Multi-Label Learning: Multi-label learning [33] is increasingly required by more and more computer vision applications including AQA. Recently, A novel MTL-AQA
dataset containing fine-grained action labels was proposed by [19]. Besides the final score label of each sequence, the action class and commentary labels are also provided for the usage of multi-label learning. In this work, instead of using extra label information from other tasks, the individual score labels from all judges and the action difficulty label are utilized by our method for multi-label learning with the expectation of a better AQA performance.

3. Approach

3.1. USDL

Pipeline Overview: For a given input video with $L$ frames $V = \{F_i\}_{i=1}^L$, we utilize a sliding window to segment it into $N$ overlapped clips where each clip contains $M$ consecutive frames. The collected clips are further sent into a backbone of Inflated 3D ConvNets (I3D) [2] followed by three fully connected layers, resulting in $N$ features as $\{f_1, f_2, ..., f_N\}$. The weights of the fully connected layers are shared among different clips. As shown in Figure 2 to deal with the extracted features, a straightforward way adopted by the most existing AQA approaches is to fuse them by average pooling or max pooling, and regress it to a final score prediction. In this work, different from those methods, we utilize a USDL framework to deal with the intrinsic ambiguity in the AQA scores, which will be detailed as below.

Score Distribution Generation: During the training phase, given a video associated with the labeled score $s$, we first generate a Gaussian function with the mean of $s$ and standard deviation of $\sigma$ as follow:

$$g(c) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(c-s)^2}{2\sigma^2}\right). \quad (1)$$

Here $\sigma$ is a hyper-parameter which serves as the level of uncertainty for assessing an action. By uniformly discretizing the score interval into a set of scores $c = [c_1, c_2, ..., c_m]$, a vector is utilized to describe the degree of each score as $g_c = [g(c_1), g(c_2), ..., g(c_m)]$. The final score distribution label $p_c = [p(c_1), p(c_2), ..., p(c_m)]$ is generated by normalizing $g_c$ as below:

$$p(c_i) = \frac{g(c_i)}{\sum_{j=1}^m g(c_j)}, \quad i = 1, 2, ..., m. \quad (2)$$

Learning from Score Distribution: In order to learn from the obtained distribution $p_c$, we map the $N$ learned features $\{f_1, f_2, ..., f_N\}$ into $N$ predicted scores as $\{s_1, s_2, ..., s_N\}$, where $s_n$ has the same size with the $p_c$. Then, the temporal average pooling is performed to $s_n = \{s_1, s_2, ..., s_N\}$ for an output vector $s'$. After the softmax activation on $s'$, we obtain the final predicted score as $s_{pre} = \{s_{pre}(c_1), s_{pre}(c_2), ..., s_{pre}(c_m)\}$. Finally, the learning loss is calculated as the Kullback-Leibler (KL) divergence between $s_{pre}$ and $p_c$:

$$KL(p_c||s_{pre}) = \sum_{i=1}^m p(c_i) \log \frac{p(c_i)}{s_{pre}(c_i)}. \quad (3)$$

Inferring from Score Distribution: During the inferring stage, we forward the input testing video into our optimized model to obtain the corresponding predicted score distribution $s_{pre}$. The final assessment is obtained by selecting the score with the max probability:

$$s_{final} = \arg \max_{c_i} \{s_{pre}(c_1), s_{pre}(c_2), ..., s_{pre}(c_m)\}. \quad (4)$$

3.2. MUSDL

In most existing works on AQA, the network is designed and optimized based on the single score supervision. However, in many Olympic games (e.g., diving, figure skiing, etc.), the final score is calculated based on multiple scores.
Learning from Multi-path Score Distributions: As shown in Figure 3 for each single-path, we use the same pipeline as our USDL approach. The fully connected layers of different paths are separately trained while the I3D backbone is shared among paths. In the training phase, suppose we have a set of scores \( \{ s_{\text{pre},k}^{\text{judge}} \}_{k=1}^{K} \) from \( K \) different judges. We first sort the scores in increasing order in order to train the sub-networks representing judges of different rigour. Following Eqn (1), we generate \( K \) Gaussian distributions as \( \{ p_{c,k}^{\text{judge}} \}_{k=1}^{K} \). Given a training video, we first feed it through the I3D backbone and obtained the distributions as different Gaussian distributions and utilize a similar strategy to train a model to predict \( DD \) by introducing a side-network branch for it during training. The predicted \( DD \) is directly used for Eqn \( (7) \) during inference. We will report the results of these two cases (\( DD \) is available or unavailable during testing) in the Experiment Section.

4. Experiment

4.1. Datasets and Experiment Settings

AQA-7 [17]: The AQA-7 dataset contains totally 1189 samples from 7 sports: 370 from single diving - 10m platform, 176 from gymnastic vault, 175 from big air skiing, 206 from big air snowboarding, 88 from synchronous diving - 3m springboard, 91 from synchronous diving - 10m platform and 83 from trampoline. We followed the setting in [17] and excluded the trampoline category in which the videos are much longer than those in the other categories [16]. There were 803 clips for training and 303 clips for testing.

MTL-AQA [19]: The MTL-AQA dataset is the currently largest dataset for AQA. There are 1412 fine-grained samples collected from 16 different events with various views in MTL-AQA. This dataset covers the events of both individual & synchronous divers, both male & female athletes, both 3m board & 10m platform settings. In this dataset, different kinds of annotations are provided to enable the study for different tasks consisting of action quality assessment, action recognition and commentary generation. Furthermore, the raw annotations of the scores from seven judges and the difficulty degree (\( DD \)) are available for each action. We followed the evaluation protocol suggested in [19] to divide the dataset into a 1059-sized training set and a 353-sized test set.
JIGSAWS [8]: Besides sport events, we further evaluate our methods on the JIGSAWS [8] dataset which contains surgical activities. There are 3 tasks as “Suturing (S),” “Needle Passing (NP)” and “Knot Tying (KT)” in this dataset. Since each video is annotated with multiple annotation scores assessing different aspects of a video (e.g., flow of operation, quality of final product, etc.) and the final score is defined as the sum of these sub-scores, we could easily extend our MUSDL based on this rule. There are stereo samples recorded by left and right cameras in this dataset, and we only used the videos from the left view due to the high similarity between the paired captures. We adopt a similar four-fold cross validation [3][16]. Since the concrete splits are not public available for [16], and the splits in [3] are specifically designed for pair-wise rank evaluation, which are not suitable for our work and [16], we simply divided the dataset into four folds randomly. Concretely, our splits are released at the project page.

Evaluation Protocols: To keep alignment with existing literatures [16], we used Spearman’s rank correlation (ranging from -1 to 1, the higher the better) to measure the performance of our methods between the ground-truth and predicted score series. Spearman’s correlation is defined as:

$$\rho = \frac{\sum_i (p_i - \bar{p})(q_i - \bar{q})}{\sqrt{\sum_i (p_i - \bar{p})^2 \sum_i (q_i - \bar{q})^2}} \quad (8)$$

Here $p$ and $q$ represent the ranking of two series respectively. Fisher’s z-value [17] is used to measure the average performance across actions.

4.2. Implementation Details

Our proposed methods were built on the Pytorch toolbox [20] and implemented on a system with the Intel(R) Xeon(R) E5-2660 v4 CPU @ 2.00Ghz. We trained our model with two Nvidia GTX 1080 Ti GPUs. We utilized the I3D model pretrained on Kinetics dataset as a feature extractor. It took action sequences containing 16 frames as inputs and output a feature with 1024 dimensions. In AQA-7 and MTL-AQA, the videos are all 103 frames long. We divided each video into ten segments under a certain rule [3]. The MLP block, containing two hidden layers $FC(256, ReLU)$ and $FC(128, ReLU)$, together with temporal pooling layer and softmax layer, built the score distribution for each video. We performed temporal pooling at feature-level on MTL-AQA and JIGSAWS for better results. Adam [14] was adopted for network optimization. In our experiments, we normalized the final total score in both datasets and seven judge scores in MTL-AQA. For the final total score, since it was a float number, we normalized it as:

$$S_{\text{normalized}} = \frac{S - S_{\text{min}}}{S_{\text{max}} - S_{\text{min}}} \times 100. \quad (9)$$

Here $S_{\text{max}}$ and $S_{\text{min}}$ indicate the maximum and minimum score in the dataset. For judge scores in MTL-AQA dataset, since these scores are inherently discrete but not an integer, so we normalized them by doubling the value of original score to get an integer. After generating normalized score $S_{\text{normalized}}$, we produced a Gaussian function with a mean of $S_{\text{normalized}}$. Note that the produced Gaussian is defined within $x \in (-\infty, \infty)$ in first. In practice, we truncated this initial distribution by the score range, and then discretized and renormalized this distribution function as described in Section 3.1.

We report the performance of the following baseline methods and different versions of our approach. Note that some of them are not evaluated on the AQA-7 dataset due to the absence of multiple scores.

- Regression: Most existing works adopted this strategy. We modified the dimension of the last fc layer in our USDL to produce a single prediction score. During the training phase, we optimized the L2 loss between predicted score and ground-truth score.
- USDL: The proposed method in Section 3.1.
- MUSDL and MUSDL*: The proposed methods in Section 3.2, which used the ground-truth and the prediction of difficulty degree during testing respectively.
- USDL-DD: During the training phase, we used scores from seven judges. According to the scoring rules of diving, the top-two and bottom-two scores will be eliminated. We summed the remaining three judge scores to obtain a new score label, and applied USDL to learn this new label. In the inference period, we multiply the predicted score with the ground-truth of difficulty degree DD to generate the final result.

4.3. Results on AQA-7 Dataset

Table 1 shows the experiment results of our methods and the comparison with other AQA approaches. The simplest yet effective regression model achieves competitive performance compared with the state-of-the-art C3D-LSTM, C3D-SVR and JRG models. The proposed USDL approach obtains significant improvement for all action classes except Sync. 10m compared to the JRG model, which also utilizes optical flow. The average correlation of our USDL approach gains improvement of 6.3% and 2.5% compared to our baseline regression model and the JRG model respectively. It is also noticed that in action class Snowboard, our proposed model surpasses state-of-the-art models by a large margin about 17%, showing the strong effectiveness of the USDL approach in AQA problems.

Study on Different Distributions: We conduct 3 parallel experiments with only the type of distribution changed.

2We took No. [0, 10, 20, 30, 40, 50, 60, 70, 80, 87] frame as the beginning frame for ten segments. See supplementary material for details.
Figure 4. Score distribution change in temporal domain. The x, y, z axis represents clip number, score and probability predicted for the certain score. The 7th, 8th clips are highlighted because the athlete fell to the ground, which leads to two salient low score prediction.

Table 1. Comparisons of action quality assessment accuracy on the AQA-7 dataset.

|                | Diving | Gym Vault | Skiing | Snowboard | Sync. 3m | Sync. 10m | Avg. Corr. |
|----------------|--------|-----------|--------|-----------|----------|-----------|------------|
| Pose+DCT [22]  | 0.5300 | –         | –      | –         | –        | –         | –          |
| ST-GCN [38]    | 0.3286 | 0.5770    | 0.1681 | 0.1234    | 0.6600   | 0.6483    | 0.4433     |
| C3D-LSTM [18]  | 0.6047 | 0.5636    | 0.4593 | 0.5029    | 0.7912   | 0.6927    | 0.6165     |
| C3D-SVR [18]   | 0.7902 | 0.6824    | 0.5209 | 0.4006    | 0.5937   | 0.9120    | 0.6937     |
| JRG [16]       | 0.7630 | 0.7358    | 0.6006 | 0.5405    | 0.9013   | 0.9254    | 0.7849     |
| Ours-Regression| 0.7438 | 0.7342    | 0.5190 | 0.5103    | 0.8915   | 0.8703    | 0.7472     |
| Ours-USDL      | 0.8099 | 0.7570    | 0.6538 | 0.7109    | 0.9166   | 0.8787    | 0.8102     |

Table 2. Study on different distributions. We use USDL approach with only the type of soft distribution changed.

|                | Diving | Gym Vault | Skiing | Snowboard | Sync. 3m | Sync. 10m | Avg. Corr. |
|----------------|--------|-----------|--------|-----------|----------|-----------|------------|
| $\chi^2$ Distribution | 0.7920 | 0.7697    | 0.6532 | 0.6905    | 0.9041   | 0.8847    | 0.8015     |
| Triangle Distribution  | 0.8147 | 0.7452    | 0.5960 | 0.6555    | 0.8987   | 0.9054    | 0.7969     |
| Gaussian Distribution       | 0.8099 | 0.7570    | 0.6538 | 0.7109    | 0.9166   | 0.8878    | 0.8102     |

The $\chi^2$ distribution is generated where the mean $s$ equals to normalized score and degree of freedom equals to one-twentieth of score range. The triangle distribution is symmetric with maximum probability at normalized score. The Gaussian distribution follows the implementation in Section 4.2. And all of the distributions are truncated discretized and renormalized to fit the range of scores. As shown in Table 2, Gaussian distribution achieves the highest average correlation while the triangle distribution performs worst. However, note that there is not a distribution that performs best in all action classes. For example, triangle distribution performs best in Diving class and $\chi^2$ distribution performs best in Gym Vault class, while the Gaussian distribution performs best in an average sense. This indicates that the ground-truth score distribution may be complicated and vary a lot among different action classes.

**Visualization of Temporal Evolution:** We choose a video in Gym Vault class to visualize the temporal evolution of score distribution in Figure 4. We obtain the score distributions of 10 clips and plot them on the graph. The video has a low ground-truth score and network gives a low score prediction. From the temporal evolution of score distribution we can see how the network gives that prediction—in the 7th and the 8th segments, the athlete fell to the ground, leading to two salient low-score prediction. The two distributions thus dominate in the final distribution, and the network finally gives the low score distribution, as expected. See supplementary material for more visualization results.

**4.4. Results on MTL-AQA Dataset**

Our method is compared with several state-of-the-arts on MTL-AQA dataset in Table 3. Still, our regression model
Figure 5. A comparison of different methods in scatter plot. The plotting points are the predictions of network with y-coordinates being the predicted scores and x-coordinates being the ground-truth scores. The ground-truth samples are plotted in dotted line.

Table 3. Comparisons of performance with existing methods on the MTL-AQA dataset.

| Method                  | Sp. Corr. |
|-------------------------|-----------|
| Pose+DCT [22]           | 0.2682    |
| C3D-SVR [18]            | 0.7716    |
| C3D-LSTM [18]           | 0.8489    |
| MSCADC-STL [19]         | 0.8472    |
| C3D-AVG-STL [19]        | 0.8960    |
| MSCADC-MTL [19]         | 0.8612    |
| C3D-AVG-MTL [19]        | 0.9044    |
| Ours-Regression         | 0.8905    |
| Ours-MUSDL              | 0.9273    |

Table 4. Ablation study on the MTL-AQA dataset. For MUSDL*, ✓✓ denote that DD is only available during the training phase, but unavailable during testing phase.

| Method         | Soft DD Judges | Sp. Corr. |
|----------------|----------------|-----------|
| Regression     |                | 0.8905    |
| USDL           | ✓              | 0.9066    |
| USDL-DD        | ✓ ✓ single    | 0.9231    |
| MUSDL*         | ✓ ✓ multiple  | 0.9158    |
| MUSDL          | ✓ ✓ single    | 0.9273    |

obtains a competitive result compared with the state-of-the-arts and our MUSDL model outperforms all the other approaches listed out. These experiment results convincingly illustrate the effectiveness of our method. It is believed that embedding the MUSDL method into a regression model is able to boost the assessment accuracy.

Ablation Study: We compare our final method with 4 other methods as shown in Table 4. The vanilla regression and USDL methods do not exploit the information from DD nor the judge scores. The architecture of USDL-DD network is the same as USDL, but the distribution label is generated from multiple judge scores. The final prediction is obtained by combining DD and predicted judge scores, which is the same as MUSDL method. The MUSDL* method adds an additional branch to the previous MUSDL method to perform a multi-task learning, i.e., training the network to predict scores from seven judges and DD at the same time. In inference stage, we combine the judge scores and DD predicted from network to obtain the final score. From the results we see that the USDL-DD outperforms the USDL by 1.7%, indicating that DD is a significant factor in diving score assessment. We believe the reason why using DD can promote better performance is that it “disentangles” the problem, making the main pipeline more specialized in video quality assessment. The MUSDL outperforms the single path method by 0.4%, which shows the fine-grained scores can further improve the performance of network.

Visualization: We first evaluate the methods in ablation study using cumulative score (CS) curves as shown in Figure 6. The cumulative prediction accuracy at the error ε is computed as $CS(\alpha) = \frac{N_{\epsilon \leq \alpha}}{N} \times 100\%$. Here $N_{\epsilon \leq \alpha}$ is the number of videos on which the prediction error $\epsilon$ is not larger than the threshold $\alpha$. The figure shows the strong
effectiveness of using DD for learning better distributions.

Then we choose the Regression, USDL and MUSDL methods and plot the scatter diagram to make a further comparison, as shown in Figure 5. The predicted scores are plotted in scatter points and the ground-truth scores are plotted in dotted line. From the results, we can see that the proposed MUSDL obtains a very satisfactory performance since the predicted points converge well to the ground-truth line.

We also conduct a case study to analyze the behavior of our network, as shown in Figure 7. The first case is a successful case where the final score predicted by MUSDL fully matches with ground-truth, and final score predicted by MUSDL* network is a little higher than ground-truth. The second case is a failure case, with the ground-truth label of the video being zero. This time both of our MUSDL network and MUSDL* network give the score far higher than the ground-truth label. The results indicate that when assessing videos with extreme scores, i.e., too far from the normal score level, the proposed method may not be able to get satisfactory results. This is because the samples with extreme scores are scarce and also because the intrinsic property of our method makes it harder to give extreme results.

4.5. Results on the JIGSAWS dataset

We finally conduct experiments on the JIGSAWS dataset for surgical activities. Because the lengths of each videos are much longer that those in AQA-7 and MTL-AQA datasets, we uniformly sampled 160 frames of each videos and divided them into 10 segments as the inputs of our models. Table 5 presents the results of our methods compared with the state-of-the-arts. Our MUSDL achieves best performance of 0.71 (S), 0.69 (NP), 0.70 (Avg. Corr.) and comparable result of 0.71 (KT), which demonstrates its effectiveness for action quality assessment.

5. Conclusion

In this paper, we have proposed a new uncertainty-aware score distribution learning (USDL) method for action quality assessment, which aims to address the inherent ambiguity in the score label. Moreover, we have devised a multi-path uncertainty-aware score distribution (MUSDL) framework to take advantage of additional fine-grained score labels. The experiments on three AQA datasets have demonstrated the effectiveness of our approach. In the future, we plan to apply our methods for instructional video analysis [28][30]. Besides, it is an interesting direction to explore the interpretability of AQA models (e.g., understanding how the network gains the score for a certain action), which is virtually important for real-world applications.

6. Acknowledgement

This work was supported in part by the National Key Research and Development Program of China under Grant 2017YFA0700802, in part by the National Natural Science Foundation of China under Grant 61822603, Grant U1813218, Grant U1713214, and Grant 61672306, in part by the Shenzhen Fundamental Research Fund (Subject Arrangement) under Grant JCYJ20170412170602564, in part by Tsinghua University Initiative Scientific Research Program, in part by National Science Foundation grant IIS-1619078, IIS-1815561, and in part by the Army Research Office ARO W911NF-16-1-0138. The authors would sincerely thank Xumin Yu, Wanhua Li, Jia-Hui Pan and Pari-tosh Parmar for their generous helps.
Figure 8. Evolution of the score distributions in temporal domain. The x, y, z axis represents clip number, score and probability predicted for the certain score. The two samples are selected from the Diving action in AQA-7 [17].

Table 6. Comparisons of different strategies to segment on the Diving action of AQA-7 [17].

| Method      | 6-seg [19] | 10-seg-s1 | 10-seg-s2 |
|-------------|------------|-----------|-----------|
| Sp. Corr.   | 0.7642     | 0.8099    | 0.7928    |

Appendix A. Different Strategies for Segmenting Videos

During our experiments, we explored three strategies to divide the videos into multiple segments. Table 6 presents the results on the Diving action of the AQA-7 dataset [17], where the length of each video clip is 103 frames. Specifically, “6-seg” denotes the scheme used in [19], which first normalized the video into 96 frames, and then divided them into 6 segments where each clip contained 16 frames³. In fact, as pointed in the recent work [16], 10 is more proper for the number of segments. Based on this, we further studied two schemes. The first used [0, 10, 20, 30, 40, 50, 60, 70, 80, 87] as the indices of beginning frames (denoted as “10-seg-s1”). Since 103/10 = 10.3, we set the stride to be 10 in most cases. And the last beginning index was set to be 87 due to the length of the video is 103 frames. The second scheme utilized [0, 9, 19, 29, 38, 48, 58, 67, 77, 87] as the indices of beginning frames (denoted as “10-seg-s2”). As shown from the results, “10-seg-s1” achieves best result among the three. Hence, we applied this scheme to the other actions in AQA-7 and MTL-AQA datasets in our paper.

Appendix B. Visualization of Temporal Evolution

In Section 4.3, we present a visualization result on the Gym_Vault action [17]. Here we further display two visualization results on the Diving action [17] in Figure 8. As it illustrates, the stage that the player enters the water plays a prominent part for action quality assessment. For example, in the bottom instance, the player causes a large splash from the 8th segment to the 10th segment. Hence, for these segments, the distributions reach the peak at the low-level scores.

References

[1] Gedas Bertasius, Hyun Soo Park, Stella X. Yu, and Jianbo Shi. Am I a baller? basketball performance assessment from first-person videos. In ICCV, pages 2196–2204, 2017.
[2] João Carreira and Andrew Zisserman. Quo vadis, action recognition? A new model and the kinetics dataset. In CVPR, pages 4724–4733, 2017.
[40] Qiang Zhang and Baoxin Li. Relative hidden markov models for video-based evaluation of motion skills in surgical training. *TPAMI*, 37(6):1206–1218, 2015.

[41] Ying Zhou, Hui Xue, and Xin Geng. Emotion distribution recognition from facial expressions. In *MM*, pages 1247–1250, 2015.

[42] Aneeq Zia and Irfan A. Essa. Automated surgical skill assessment in RMIS training. *Int. J. Comput. Assist. Radiol. Surg.*, 13(5):731–739, 2018.

[43] Aneeq Zia, Yachna Sharma, Vinay Bettadapura, Eric L. Sarin, and Irfan A. Essa. Video and accelerometer-based motion analysis for automated surgical skills assessment. *Int. J. Comput. Assist. Radiol. Surg.*, 13(3):443–455, 2018.

[44] Aneeq Zia, Yachna Sharma, Vinay Bettadapura, Eric L. Sarin, Thomas Ploetz, Mark A. Clements, and Irfan A. Essa. Automated video-based assessment of surgical skills for training and evaluation in medical schools. *Int. J. Comput. Assist. Radiol. Surg.*, 11(9):1623–1636, 2016.