FALCONS: FAst Learner-grader for CONtorted poses in Sports

Mahdiar Nekoui
University of Alberta
nekoui@ualberta.ca

Fidel Omar Tito Cruz
Universidad Nacional de Ingeniería
ftitoc@uni.pe

Li Cheng
University of Alberta
lcheng5@ualberta.ca

Abstract

Isn't it about time to help judges with the challenging task of evaluating athletes' performances in sports with extreme poses? To tackle this problem and inspired by human judges' grading schema, we propose a virtual refereeing network to evaluate the execution of a diving performance. This assessment would be based on visual clues as well as the body joints sequence of the action video. In order to cover the unusual body contortions in such scenarios, we present ExPose: annotated dataset of Extreme Poses. We further introduce a simple yet effective module to assess the difficulty of the performance based on the extracted joints sequence. Finally, the overall score of the performance would be reported as the multiplication of the execution and difficulty scores. The results demonstrate our proposed lightweight network not only achieves state-of-the-art results compared to previous studies in diving but also shows acceptable generalization to other contortive sports.

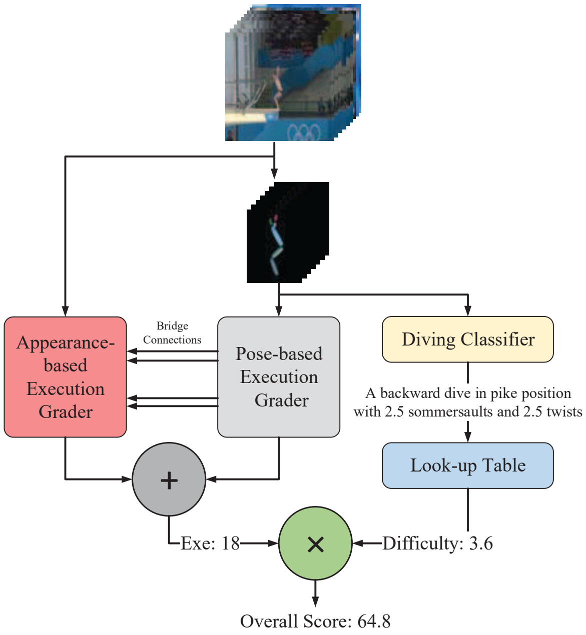

Appearance-based Execution Grader
Bridge Connections
Pose-based Execution Grader
Diving Classifier
A backward dive in pike position with 2.5 sommersaults and 2.5 twists
Look-up Table
+
Exe: 18
×
Difficulty: 3.6
Overall Score: 64.8

Figure 1: Overall pipeline of FALCONS.

1. Introduction

Sports is the language of joy and unity. Sporting events are usually among the top most-watched televised broadcasts [3]. Fairness in evaluation is of utmost importance to both the competitors and spectators, hence the need for a structured means to evaluate the athletes and determine the winner. In recent years, the advent of technology has brought more just and agile refereeing to soccer games by the introduction of video assistant referees (VAR). However, some popular fields like diving, gymnastics, etc., are still suffering from the inefficiency of human-based judging systems. For example, in a typical diving contest, 7 judges score the performance of each athlete. These judges should be from different nationalities to that of the contestant, limiting the qualified choices. Furthermore, although the international swimming federation (FINA) has prohibited the judges from looking at the replays to make the grading procedure faster, it takes about 40 seconds to report the score for each performance that itself takes only about 4 seconds. Considering all these issues, it would be a great help to introduce an automatic grading system to score the athletes in a faster and more accurate manner by reducing human intervention. The task of evaluating how an action has been performed and assigning a grade to the performer is known as Action Quality Assessment (AQA).

To date, the literature of AQA has been dominated by networks that have tried to regress the overall score of each athlete as the only label that needs to be predicted [11, 13, 14, 17]. However, each performance should be considered as a complex action in which not only the quality of the execution but also the difficulty of the task contributes to the final score. Nevertheless, the judges are only responsible for awarding the execution score and the difficulty score would be determined based on a predefined official benchmark released by the corresponding federation of that sport. For example, in a diving contest, the difficulty score of each performance would be awarded based on the type of rotation, position of diving, number of sommersaults, and the number of twists according to FINA diffi-

2160-7516/20/$31.00 ©2020 IEEE
DOI 10.1109/CVPRW50498.2020.00458

culty look-up table [4]. On the other hand, the judges award the execution score based on the appearance-based features of the flight (*e.g*. smoothness and aesthetic pleasure of the flight, and amount of splash), and also pose-based features (*e.g*. angle of entry to the water).

Existing literature regarding automatic graders take into account either the pose-based or appearance-based features to regress the final score of an action, hence suffering from the limited performance [14, 17, 8, 13, 12]. In this paper, we decompose the overall score into execution and difficulty. For the former, inspired by what a human judge does, we propose a virtual refereeing system that considers both the pose-based and appearance-based features as the contributors to the execution score. As for the latter, we introduce a difficulty extractor module that classifies the task based on the sequence of body joint arrangements throughout the performance. Consequently, the difficulty score would be determined by feeding the classes (*e.g*. type of rotation and etc.) into the difficulty look-up table. The overall pipeline of our work is depicted in Fig.1.

The extraction of pose features has its own challenges. In most of the sports video footage, not only the athlete but also the camera is moving fast to track the athlete during the performance, causing motion blurriness of the image frames. Furthermore, the unusual configuration of the athletes' poses is not covered by the existing datasets [9, 1, 7]. These facts have limited the performance of the pose estimation networks in such cases. To address this problem we introduce *ExPose*, a dataset of extreme poses which includes 4000 annotated blurry images in extreme body contortion scenarios. In order to regress the scores based on the extracted joint sequences (taken from the pose estimation module trained on our dataset), we develop the well-known Spatial Temporal Graph Convolutional Networks (ST-GCN) [19] to also learn dependencies between unconnected joints. ST-GCN extends graph convolutional networks to simultaneously capture spatial and temporal features for action recognition. However, it is only able to learn correlations between the directly connected joints. This fact may affect the performance of the automatic grader in which the symmetry of different parts of the body should contribute to the execution score. To address this issue, we introduce the idea of virtual super-joints. Each super-joint is simply the average of its constituting joints' location and may have connections with other super-joints.

For spatio-temporal appearance features extraction, we divide the whole task video into T subtasks, each consisting of N consecutive frames. In order to learn global spatio-temporal features, we follow [5] that applies 2D spatial convolution filters followed by 1D temporal ones on each subtask. Finally, the temporal dynamics between the subtasks should be encoded to an execution score. However, as the amount of splash plays a greater role than other appearance features in the execution assessment, not all of the subtasks should have the same contribution to the score. To this end, we propose a bridge-connection module that fuses the feature sequence of each subtask with a contribution weight. This module takes the average of confidence score of pose estimation in each subtask and maps it to a weight for each. Simply put, as the performer is under the water in splash capturing frames, the lower the average confidence score of the subtask, the higher its contribution to the Exe score.

Finally, the weighted sum of the score of the joint-based and appearance-based graders generates the execution score (see Fig.1). The overall score would then be calculated by multiplication of the extracted execution and difficulty scores. To validate the effectiveness of our method we conducted experiments on existing datasets and demonstrate that our method not only achieves state-of-the-art results in diving grading but also shows acceptable generalization to other fields. The main contributions of this paper are summarized as follows:

- We introduce *ExPose*, a dataset of blurry images with annotated joints in extreme contortions to facilitate extreme pose estimation in sports analytics.
- We propose a difficulty extractor module that leverages the pose sequence to determine the type and details of the dive and award a difficulty score based on the FINA look-up table. This module doesn't need to be trained in advance and achieves state-of-the-art results.
- Inspired by human judges grading schema, we present a novel virtual refereeing system that regresses the execution score by leveraging both appearance and pose-based features. The results demonstrate the superior performance of the proposed network over previous works with promising generalization to other sports.

2. Related work

Pose-based AQA

Due to the challenges of estimating pose in extreme configurations of athletes' body and blurriness of image frames, pose-based AQA has been largely unexplored. Pirsiavash *et al*. [14] presented the first model for assessing the quality of a sports action based on pose features. They trained a linear SVR on DCT frequency coefficients of the action pose features to regress the final score. Recently, Pan *et al*. [11] decoupled pose features to body parts kinetics and joints coordination, and provided the score by capturing graph-based joint relations. However, these networks neglect the collaborative role of appearance-based with pose-based clues in score awarding. Furthermore, they rely on the overall score as the only label that needs to be predicted. In addition, their pose estimation modules have been trained on regular existing pose datasets that don't cover unusual contorted

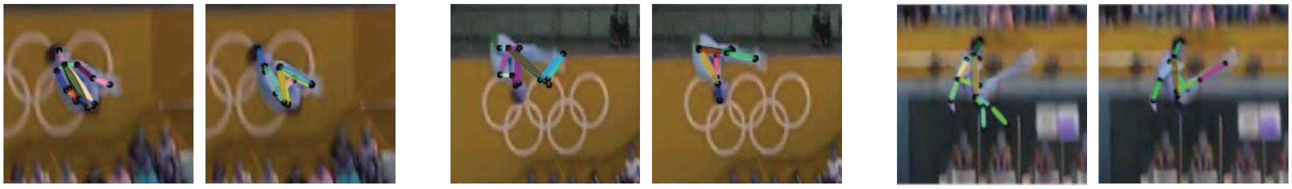


Figure 2: The pose estimation network should be trained on a dataset which covers blurry images with extreme configuration of joints. The left side picture of each column is the result of training HRNet on MPII [1] and the right side one corresponds to the results of training the network on our ExPose dataset. As evident, the network trained on our data set better identifies body's true joint locations.

pose configurations of Olympics divers. The above reasons generally lead to the underperformance of such networks.

Appearance-based AQA

On the other end of the spectrum, most studies in the realm of AQA rely exclusively on visual clues of the action in order to regress the score [8, 13, 18, 12]. Parmar and Morris [13] use C3Ds [16] (inflated counterpart of 2D ConvNet) to capture visual spatio-temporal features of the action and feed them to a regression framework (SVR, LSTM, and LSTM-SVR) to provide the final score. Li *et al*. [8] segment videos to multiple fragments and feed them to parallel C3Ds, followed by some 2D convolution layers. They discuss that these fragmented features would have more distinctive power to differentiate a good performance from a bad performance. However, the heavy computation cost of using C3Ds, as well as neglecting the pose features and predicting only the overall score has affected the performance of these models. To address the last issue, Parmar and Morris [12] recently proposed a multi-task approach that jointly learns commentary, action class details, and AQA overall score based on the appearance features extracted from C3Ds. Nevertheless, the first two issues still persist in the latter study.

3. Approach

In the following, we present our proposed network consisting of two modules. The first module, virtual referee, evaluates the performance execution based on both of the appearance and pose features. The second module, difficulty assessor, evaluates the difficulty based on the joints sequence of the action.

3.1. Virtual Referee

Fig. 3 demonstrates our virtual referee pipeline which is comprised of pose-based and appearance-based assessors. In what follows we elloborate more on the details of the two assessors.

Pose-based execution assessor

We use HRNet pose estimation network [15] to extract the pose features of extreme actions. The network has been trained on our ExPose dataset to be able to handle blurry images with extreme body contortions. As depicted in Fig. 2, the performance of training the network on regular datasets is not acceptable.

ExPose dataset contains 3000 diving and 1000 gymnastic vault images together with their annotations. The diving images are obtained from four different individual diving events recorded from side-view on two types of boards. The springboard images are obtained from men's 3m final of the 2019 world series and the platform images are taken from men's 10m platform finals of the 2016 European aquatics championships in London, 2018 youth Olympic games in Buenos Aires, and 2019 world series in Sagamihara. The gym vault images are obtained from three different events; men's and women's final of the 2018 Doha world championships and women's final of the 2018 Glasgow European competitions. All the original videos of the both diving and gym vault images are sourced from YouTube. For consistency with existing datasets, the annotations have been provided in MPII dataset format, considering 16 joints for the human body.

The extracted joint sequence should be fed into an action regressor that is able to learn the spatio-temporal skeleton-based features of the action. To this end, we use ST-GCN [19] that considers the joints of a skeleton as the nodes of a graph. Spatial edges of the graph connect the structurally neighboring joints and temporal edges connect the same node in consecutive frames. However, this network is not able to capture the dependencies between joints that are unconnected in the predefined skeleton graph.

In order to capture the non-local information (for unconnected joints), we introduce three levels of virtual super-joints. Each super-joint represents the average location of its constituting real joints. In the first level, the super-joints consist of the right leg, left leg, right hand, left hand, waist, and head. As a result, we would be able to capture wrists-

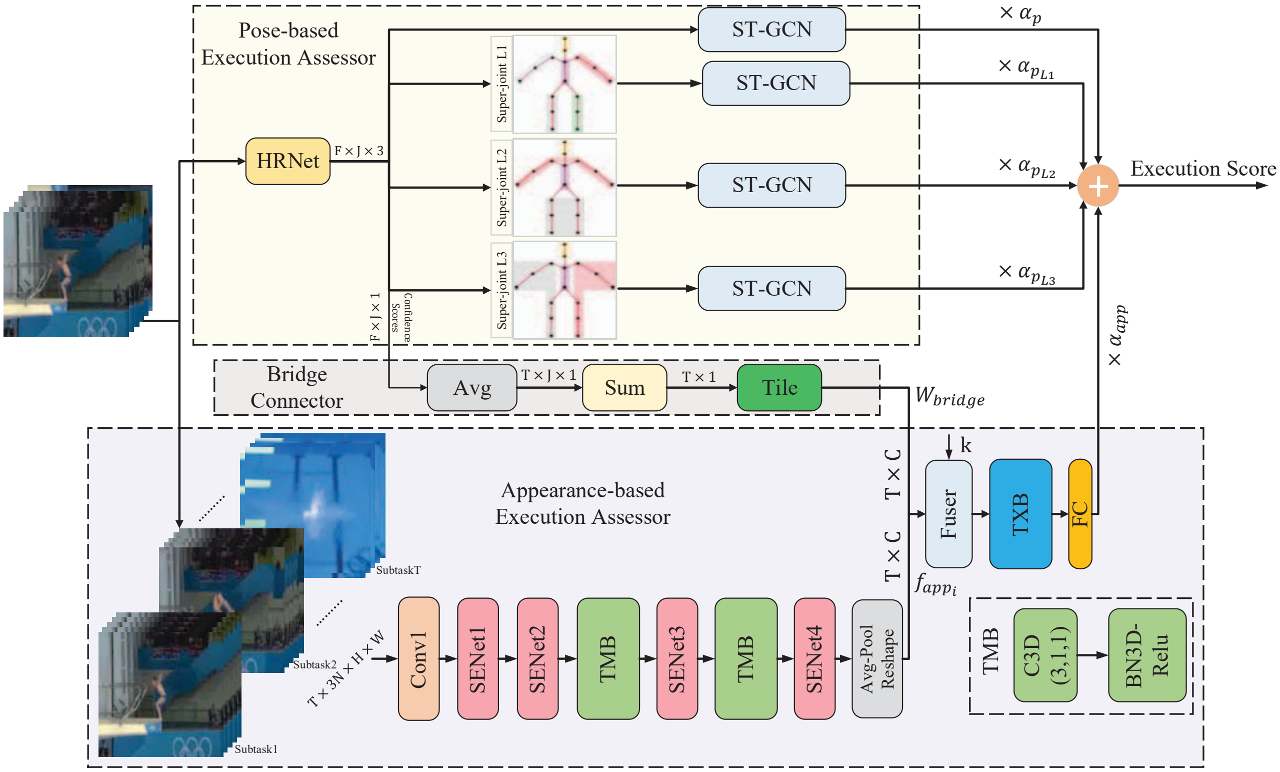


Figure 3: Overview of our virtual referee pipeline. This network generates the execution score of the performer based on appearance features as well as pose ones. The bridge connector module links these two to increase the contribution weight for the appearance features with the most importance (splash).

shoulders and hips-ankles relationships. The second super-joint level captures dependencies between the upper body and the lower body. Finally, the third super-joint level extracts the symmetricity between the left and the right side of the body. The composition of super-joints is visualized in the upper part of Fig. 3.

It goes without saying that not only the local relations between connected joints but also richer dependencies like symmetricity between different parts of the body should verify how well the action was performed. Thus, the features of the fixed predefined skeleton body (consisting of all 16 body joints) as well as non-local features of virtual super-joints levels are fed into four parallel ST-GCNs to assess the coordination between the joints and between the body parts (see the upper part of Fig. 3). The output value of each ST-GCN module is multiplied by its own contribution weight to account for the superior importance of some features over others. For example, intuitively, the contribution weight of the third level of super-joints score should be larger as it is responsible for the balance of the body during the flight.

Appearance-based execution assessor

The proposed network should not only be effective but also have a lightweight configuration to act as fast as possible in both training and testing phases. To this end, we follow StNet [5] structure that instead of using stacked heavyweight C3Ds, decomposes 3D convolutions into 2D spatial convolutions followed by 1D temporal ones. First, we divide the whole task into T subtasks, each consisting of N frames (in this case T=17, and N=6). Each subtask is fed to 2D convolution layers to extract the local spatio-temporal information. The Conv1 module followed by the 2 SENet [6] layers are responsible to do so (see the lower part of Fig. 3). The extracted local features are fed to a Temporal Modelling Block (TMB), introduced by [5], to get the temporal features across each subtask. The TMB block is consisted of a 3D convolutional layer followed by a BN3d-ReLU. As the spatial information is already captured by SENets, the spatial filter size of the 3D convolution for TMB is set to 1. In order to get deeper correlations, the result of the TMB module is further passed to a stack of SENet-TMB-SENet.

The extracted features (f_{app_i}) would contain local spatio-temporal information of each subtask as well as temporal dynamics across N frames of each subtask. In the next step, we should capture the temporal information between the T subtasks. However, there are some visual clues that should contribute to the execution score more than the others. Performing a rip entry into the water making the least amount of splash is more important than other appearance-based clues like the smoothness of the flight. In order to address the aforementioned concern, we introduce a bridge connection module to increase the contribution weight of the final subtasks which is where the diver makes an entry into the water. As discussed before we use HRNet module to detect the joints in each frame. This module not only provides the joints' locations but also gives the overall confidence of estimation for each joint. For example, in the last frames in which the athlete is under the water, the very distorted visibility of the human body in such frames causes a drastic drop in the confidence score of its joints estimation. Inspired by this fact, we introduce a link module to set the weights of subtasks based on the confidence score of pose estimation in their frames (see the middle part of Fig. 3). Given the confidence scores of all frames, we first take the average between the confidence scores of each subtask. Thus, considering the output of HRNet as a $F \times J \times 1$ tensor (where $F = T \times N$ is the number of frames and J is the number of joints), the output of the average module would have size $T \times J$. In the next step, the confidence scores of all joints in each subtask have been summed to form a $T \times 1$ vector and normalized to $(0-1)$ interval. Finally, the vector is tiled to have the same size as that of the extracted appearance features ($T \times C$). The Fuser module takes these two (f_{app_i} and W_{bridge}) as well a scale value (k) to set the contribution of each subtask using the below formula.

$$f_{app_o} = f_{app_i} \odot (1 - W_{bridge}(1-k)) \tag{1}$$

As a result, the lower W_{bridge} a subtask has, the higher contribution to the score regression it would have. Finally, to capture the temporal information between the subtasks, the resulting f_{app_o} is fed to a Temporal Xception Block (TXB). This block, introduced by [5], decomposes the temporal dynamics of the extracted feature sequence into a 1D temporal-wise and a 1D channel-wise convolutions. Deploying this strategy instead of averaging between the features of the subtasks has boosted the results in action classification tasks. Finally, the resulting tensor of the TXB module has been fed to a fully connected layer to regress a number as the appearance based execution score.

In the end, as evident in Fig. 3, the virtual referee calculates the weighted sum of the appearance-based execution score and the pose-based ones (including regressed scores of virtual pose levels) to award the final execution score:

$$\begin{aligned} S_{Exe_f} = \alpha_{app} S_{Exe_{app}} + \alpha_p S_{Exe_p} + \alpha_{p_{L1}} S_{Exe_{p_{L1}}} \\ + \alpha_{p_{L2}} S_{Exe_{p_{L2}}} + \alpha_{p_{L3}} S_{Exe_{p_{L3}}} \end{aligned} \tag{2}$$

where α_{app}, α_p, and $\alpha_{p_{Li}}$ represent the contribution weight of appearance-based, predefined skeleton pose-based, and virtual-pose levels execution assessment.

3.2. Difficulty assessor

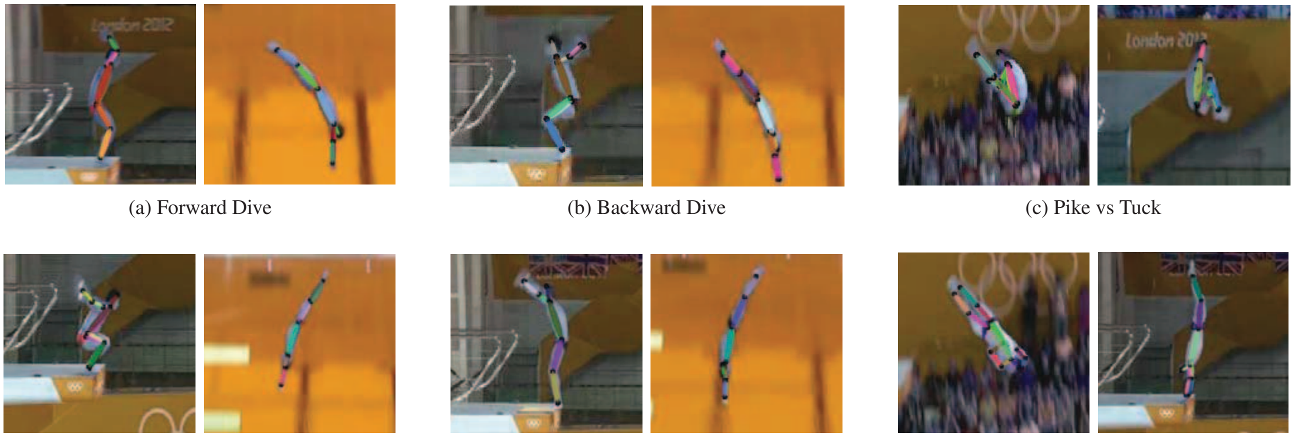
Given the joint sequences extracted from the HRNet module and the direction of filming (west-side or east-side camera), our proposed difficulty assessor classifies the performed dive and provides the difficulty score based on the FINA look-up table.

In terms of the rotation type, a performance can be classified into four different groups: forward, reverse, backward, and inward. The group can be determined based on the joints position during frames that capture take-off and entry into the water. For example, when the camera is located in the east-side of the platform and the athlete faces the front of the board (forward or reverse; Fig. 4a,4d), x_{knee} would be greater than both x_{hip} and x_{ankle} (the origin of the coordinates is located at the top left of the picture). On the other hand, when the athlete takes off with his (her) body back to the water (backward or inward; Fig. 4b,4e), x_{knee} would be less than both x_{hip} and x_{ankle}. Another metric that helps determine the group is x_{ankle}'s value with respect to other joints' positions during frames that capture entry into the water. As it can be seen in Fig. 4a,4b,4d,4e, the joints position in the take-off frame and entry one together distinct a rotation type from another.

For getting the position of the dive, the module looks for a specific pattern among all of the frames. In a pike position, the body is bent at the waist but the legs should remain straight (see Fig. 4c). On the other hand, in a tuck position the knees should be pulled tightly to the chest. If a position is neither a pike nor a tuck it would be considered as a free dive. Thus, based on angle between the lower leg and thigh we would be able to determine the position.

In a regular dive we only have the sideview profile of the performance. However, in a twisted dive the performer rotates its body around the vertical axis, exposing the full-body profile during the performance (see Fig. 4f). In order to discern a sideview profile from a frontview one, we have set a threshold for captured shoulder-width of the athlete during the performance. The #twists can be determined based on number of the switches from a side-view to a front-view profile and vice versa.

In order to get the #sommersaults, the module monitors the relative positions of the thorax and the pelvis joints along the y-axis. In a normal configuration of the body, the thorax should be located in a higher position than the pelvis.



(a) Forward Dive

(b) Backward Dive

(c) Pike vs Tuck

(d) Reverse

(e) Inward Dive

(f) Twisted and Armstand

Figure 4: The type of the rotation, the position, and #twists can be determined based on joint sequences. In addition, based on the the configuration of pelvis and thorax, the #sommersaults and first stand position can be determined. The performance of pose estimation should be acceptable to have a better diving classifier and respectively a better difficulty assessor.

However, this configuration changes as the body is rotated around the horizontal axis by performing a sommersault. The module counts the number of configuration switches to provide the #sommersaults. Furthermore, as evident in Fig. 4f, we can also distinguish an armstand dive from the regular one by getting the configuration of the pelvis and the thorax during the take-off frame.

Finally, the difficulty of the execution would be assessed based on the type of rotation, position, #twists, #sommersaults, and whether it was performed in an armstand position or not, according to FINA difficulty look-up table. The resulting difficulty score would be multiplied by the execution score provided by the virtual referee and the overall score determines the ranking of the competitor.

4. Experiments

Implementation Details

The weights of the all SENet blocks of the appearance-based assessor module are initialized using the ImageNet pretraied weights. As the first convolution layer takes $3N$ channels instead of 3, the same procedure as [2] have been taken to inflate the ImageNet weights.

The network has been trained for 200 epochs with a learning rate of 0.000075 using the SGD optimizer with the momentum of 0.95 and a batch size of 10. Mean squared error has been used as the loss function of the regression. The network is trained and tested on UNLV-Diving dataset [13]. The videos of the dataset are originally normalized to 103 frames. We ignore the last frame and feed the rest 102 to the network. The number of subtasks (T) is set to 17 and each one is consisted of 6 frames (N). 300 videos of the dataset are dedicated to training and the rest 70 are utilized for testing.

	MTL-AQA	**Ours**
Armstand?	98.65	**100.00**
Rotation Type	97.30	**97.57**
Position	95.14	**97.84**
#Sommersaults	95.68	**98.92**
#Twists	94.05	**94.32**

Table 1: The results of diving detailed classification on UNLV-Diving dataset [13].

4.1. Quantitative Results

Diving Classification

To the best of our knowledge, there are only two studies in detailed diving classification [10, 12]. Table 1 shows our results compared to MTL-AQA [12] which performs better than the other one. Aside from outperforming the baseline study, the proposed diving classifier runs online without the need of being trained in advance. It takes the extracted pose sequence from the HRNet module and outputs the detailed classification of the dive as well as the difficulty score. On the other end of the spectrum, the MTL-AQA network uses C3Ds that leads to high computational cost and a huge impact on both training and testing phases speed.

Although this paper is mostly geared towards diving routines, a similar procedure can be taken to assess the performance difficulty of other sports. For example, the difficulty score of a trampoline routine would be awarded based on

Method	Spearman's Corr.
Pose-DCT-SVR [14]	53.00
Joint Relation Graphs [11]	76.30
C3D-SVR (best performing in [13])	78.00
MSCADC-STL [12]	79.79
Li *et al*. [8]	80.09
C3D-AVG-STL [12]	83.83
Ours (FALCONS)	**84.53**

Table 2: The results of predicting the overall score of the athletes. Our network outperforms the previous baselines. All networks are trained and tested on UNLV-Diving dataset [13].

Ablated Model	Spearman's Corr.
Bridge Blocking	69.11
Virtual Pose Levels Removal	76.68
Only Appearance-based	81.75
Only Pose-based	50.04

Table 3: The results of systematically removing the components of the network to study their effectiveness in the final results.

the #twists and #sommersaults performed in the whole task. As another example, the difficulty of a gymnastic vault performance is assessed based on the number of turns (twists), the position of flight phase (tuck, pike, or stretched), and type of the approach towards the handspring (backward or inward).

Overall Score Assessment

In order to be consistent with existing literature, we have used Spearman's Rank correlation as the evaluation metric of our network. We evaluate our model on UNLV-Diving dataset [13]. As evident in Table 2, our proposed model achieves superior performance than prior methods. The parameters α_{app}, α_p, $\alpha_{p_{L1}}$, $\alpha_{p_{L2}}$, $\alpha_{p_{L3}}$, and k have been set to 0.9, 0.01, 0.01, 0.01, 0.07, and 0.9 respectively. It should be noted that [12] also proposes a multi-task approach in which they augmented the original dataset with captioning to jointly learn the overall AQA score, detailed diving classification, and commentary. This implementation resulted in Spearman's Rank correlation coefficient of 88.08. However, as our method is trained on the original dataset without using excessive information of captioning we have compared our results with their single-task approach.

We also conducted an extensive ablative study to evaluate the components of our network. In the first set of experiments, we blocked the bridge connector by setting the k scale of the Fuser to 1. As a result, all appearance features would have the same contribution to the execution assessment procedure. In order to study the importance of non-local pose features, we removed the contribution of the virtual pose levels by setting their contribution weights ($\alpha_{p_{L1}}$, $\alpha_{p_{L2}}$, $\alpha_{p_{L3}}$) to 0. We further set α_p to 0, relying only on appearance features for execution scoring. Finally, we set the α_{app} to 0 to evaluate the Only Pose-based execution assessor. In this experiment, α_p, $\alpha_{p_{L1}}$, $\alpha_{p_{L2}}$, and $\alpha_{p_{L3}}$ have been set to 0.1, 0.1, 0.1, 0.7 respectively to have the same scale as the original method. The results of Table 3 show how each component contributes to the effectiveness of our final model.

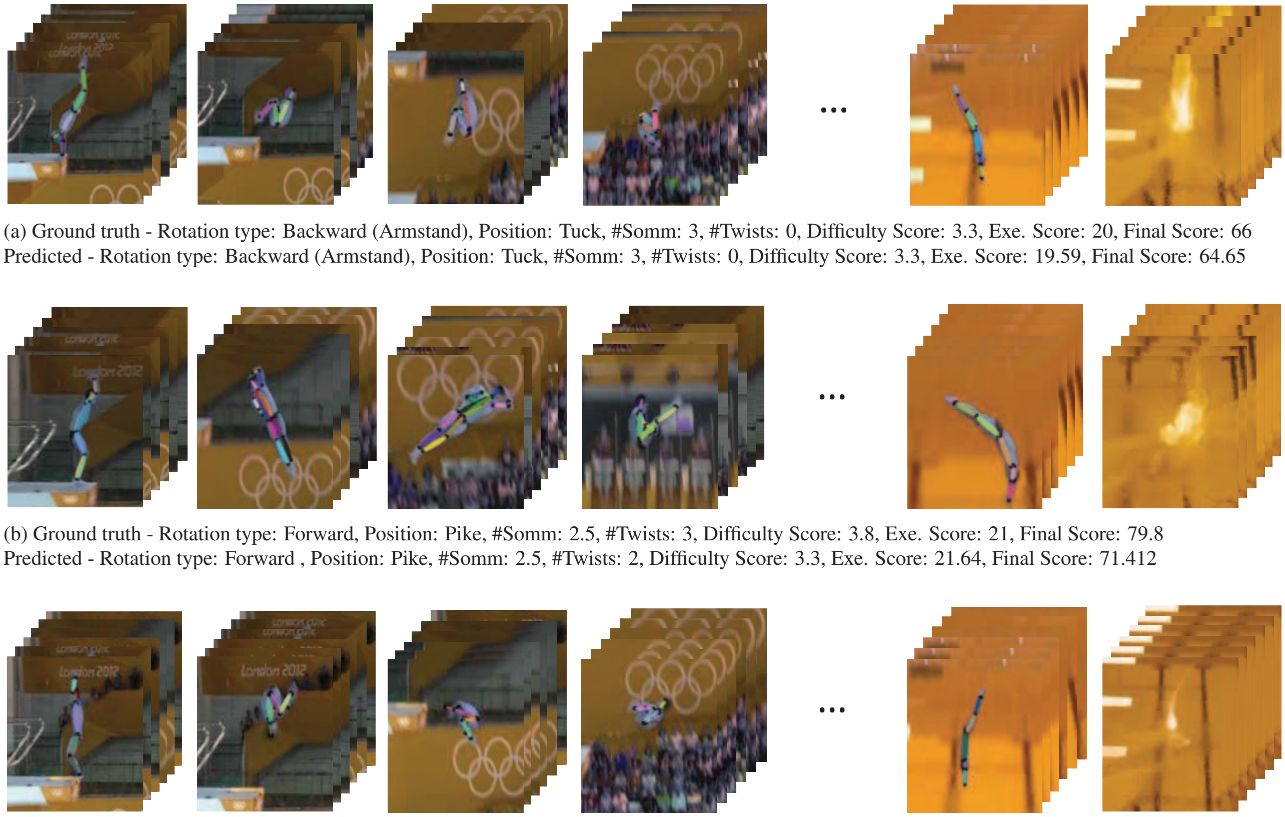
Generalization to Other Fields

Here we present the effectiveness of our method tested on an unseen event from another sport. To this end, we use our pretrained network on the UNLV-Diving dataset and fit a linear regressor on top of the resulting tensor of Fuser module to test unseen gymnastic vault routines. We freeze all other weights of the network to be unbiased around the gymnastic vault performances. As the weight contribution of the appearance features of each subtask is equally important, we blocked the bridge connector by setting the scale (k) to 1. All other parameters have remained the same as before. As we have not provided the framework for assessing the difficulty of the routine in a gymnastic vault performance (classifier + lookup table), we resort to the evaluation of our virtual refereeing system that provides the execution score.

We trained and tested the modified network on the gymnastic vault videos of AQA dataset [13]. The resulted Spearmann's Rank correlation for this task is 27.11. At the first glance it seems that the network is underperforming. However, it should be noted that the ground truth execution scores awarded by the judges are close to each other. In such condition, the difficulty score plays a distinctive role in ranking the athletes. Thus, having an effective difficulty assessor contributes to higher Spearmann's Rank correlation. The promising results of our network on the challenging task of assessing an unseen event from gymnastic vault show that the proposed network can be generalized to be used in other sports.

4.2. Qualitative Results

Fig. 5 presents some qualitative results for classifying the dive as well as assessing it in terms of execution and difficulty. The visual clues like amount of splash as well as joint configuration during the performance, contribute to the execution score.



(a) Ground truth - Rotation type: Backward (Armstand), Position: Tuck, #Somm: 3, #Twists: 0, Difficulty Score: 3.3, Exe. Score: 20, Final Score: 66
Predicted - Rotation type: Backward (Armstand), Position: Tuck, #Somm: 3, #Twists: 0, Difficulty Score: 3.3, Exe. Score: 19.59, Final Score: 64.65

(b) Ground truth - Rotation type: Forward, Position: Pike, #Somm: 2.5, #Twists: 3, Difficulty Score: 3.8, Exe. Score: 21, Final Score: 79.8
Predicted - Rotation type: Forward , Position: Pike, #Somm: 2.5, #Twists: 2, Difficulty Score: 3.3, Exe. Score: 21.64, Final Score: 71.412

(c) Ground truth - Rotation type: Inward, Position: Tuck, #Somm: 4.5, #Twists: 0, Difficulty Score: 4.1, Exe. Score: 23.5, Final Score: 96.35
Predicted - Rotation type: Inward , Position: Tuck, #Somm: 4.5, #Twists: 0, Difficulty Score: 4.1, Exe. Score: 24.07, Final Score: 98.687

Figure 5: Qualitative results of our proposed method.

5. Conclusion

We present FALCONS, an engine of grading Olympic diving athletes, based on execution and difficulty assessors. Similar to what human judges do, the execution evaluation is based on both visual and pose features of the action. To handle the estimation of pose in such extreme body configurations, we introduce the *ExPose* dataset. By introducing the notion of virtual super-joints, we augment the local correlations between connected joints with non-local joint dependencies of the action. The extracted pose sequences are also utilized by the bridge connector module to increase the contribution of the splash scene among other appearance clues. For extracting the difficulty of the action we propose a simple assessor that works on the basis of pose features. Finally, the overall score is provided by the multiplication of the execution and difficulty scores. The results show state-of-the-art performance compared to previous studies as well as acceptable generalization to unseen scenes from other sports.

6. Acknowledgments

This work is partly supported by the NSERC Discovery Grant (No. RGPIN-2019-04575), and the UAHJIC grant (No. 201902).

References

[1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition*, pages 3686–3693, 2014.

[2] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[3] International Olympic Committee. London 2012 Olympic Games: Global Broadcast Report. `https://stillmed.olympic.org/Documents/IOC_Marketing/Broadcasting/London_2012_Global_%20Broadcast_Report.pdf`.

[4] International Swimming Federation. Fina diving rules. http://www.fina.org/sites/default/files/2017-2021_diving_16032018.pdf.

[5] Dongliang He, Zhichao Zhou, Chuang Gan, Fu Li, Xiao Liu, Yandong Li, Limin Wang, and Shilei Wen. Stnet: Local and global spatial-temporal modeling for action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8401–8408, 2019.

[6] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[7] Sam Johnson and Mark Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *Proceedings of the British Machine Vision Conference*, 2010. doi:10.5244/C.24.12.

[8] Yongjun Li, Xiujuan Chai, and Xilin Chen. End-to-end learning for action quality assessment. In Richang Hong, Wen-Huang Cheng, Toshihiko Yamasaki, Meng Wang, and Chong-Wah Ngo, editors, *Advances in Multimedia Information Processing – PCM 2018*, pages 125–134, Cham, 2018. Springer International Publishing.

[9] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[10] Aiden Nibali, Zhen He, Stuart Morgan, and Daniel Greenwood. Extraction and classification of diving clips from continuous video footage. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 38–48, 2017.

[11] Jia-Hui Pan, Jibin Gao, and Wei-Shi Zheng. Action assessment by joint relation graphs. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[12] Paritosh Parmar and Brendan Tran Morris. What and how well you performed? a multitask learning approach to action quality assessment. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[13] Paritosh Parmar and Brendan Tran Morris. Learning to score olympic events. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017.

[14] Hamed Pirsiavash, Carl Vondrick, and Antonio Torralba. Assessing the quality of actions. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 556–571, Cham, 2014. Springer International Publishing.

[15] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, 2019.

[16] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015.

[17] Vinay Venkataraman, Ioannis Vlachos, and Pavan Turaga. Dynamical regularity for action analysis. In Mark W. Jones Xianghua Xie and Gary K. L. Tam, editors, *Proceedings of the British Machine Vision Conference (BMVC)*, pages 67.1–67.12. BMVA Press, September 2015.

[18] Chengming Xu, Yanwei Fu, Bing Zhang, Zitian Chen, Yu-Gang Jiang, and Xiangyang Xue. Learning to score figure skating sport videos. *IEEE Transactions on Circuits and Systems for Video Technology*, 2019.

[19] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition, 2018.