Video-based Skill Assessment for Golf: Estimating Golf Handicap

Christian Keilstrup Ingwersen
cin@trackman.com
TrackMan & Technical University of Denmark
Vedbaek, Denmark

Artur Xarles
arturxe@gmail.com
Universitat de Barcelona & Computer Vision Center
Barcelona, Spain

Albert Clapés
aclapes@ub.edu
Universitat de Barcelona & Computer Vision Center
Barcelona, Spain

Meysam Madadi
mmadadi@ub.edu
Universitat de Barcelona & Computer Vision Center
Barcelona, Spain

Janus Nørtoft Jensen
jnje@dtu.dk
Technical University of Denmark
Lyngby, Denmark

Morten Rieger Hannemose
mohan@dtu.dk
Technical University of Denmark
Lyngby, Denmark

Anders Bjorholm Dahl
abda@dtu.dk
Technical University of Denmark
Lyngby, Denmark

Sergio Escalera
sescalera@ub.edu
Universitat de Barcelona & Computer Vision Center & Aalborg University
Barcelona, Spain

ABSTRACT

Automated skill assessment in sports using video-based analysis holds great potential for revolutionizing coaching methodologies. This paper focuses on the problem of skill determination in golfers by leveraging deep learning models applied to a large database of video recordings of golf swings. We investigate different regression, ranking and classification based methods and compare to a simple baseline approach. The performance is evaluated using mean squared error (MSE) as well as computing the percentages of correctly ranked pairs based on the Kendall correlation. Our results demonstrate an improvement over the baseline, with a 35% lower mean squared error and 68% correctly ranked pairs. However, achieving fine-grained skill assessment remains challenging. This work contributes to the development of AI-driven coaching systems and advances the understanding of video-based skill determination in the context of golf.

CCS CONCEPTS

• **Computing methodologies** → **Activity recognition and understanding**; **Video summarization**.

KEYWORDS

datasets, neural networks, action quality assessment, golf, action understanding

MMSports '23, October 29, 2023, Ottawa, ON, Canada

ACM ISBN 979-8-4007-0269-3/23/10…$15.00
https://doi.org/10.1145/3606038.3616150

ACM Reference Format:
Christian Keilstrup Ingwersen, Artur Xarles, Albert Clapés, Meysam Madadi, Janus Nørtoft Jensen, Morten Rieger Hannemose, Anders Bjorholm Dahl, and Sergio Escalera. 2023. Video-based Skill Assessment for Golf: Estimating Golf Handicap. In *Proceedings of the 6th International Workshop on Multimedia Content Analysis in Sports (MMSports '23), October 29, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3606038.3616150

1 INTRODUCTION

As Malcolm Gladwell famously stated, it is often said that it takes approximately 10,000 hours of deliberate practice to master a skill [11]. However, a crucial question arises: what type of practice can expedite the journey towards skill mastery? The answer to this question heavily relies on an individual's current skill level.

Automated sports coaching systems have emerged as a potential game-changer in how athletes learn and improve their skills. These systems hold the promise of providing personalized feedback and guidance, revolutionizing the coaching landscape. However, a significant challenge in developing such systems lies in accurately assessing an athlete's current skill level based solely on video input. In this paper, we tackle the problem of skill determination of golfers by leveraging a deep learning model applied to video recordings of their swings. While assessments in other sports like diving [21, 20, 22] are based on clear criteria and the current performance, golf presents unique challenges. Defining a metric that measures the quality of a single golf swing is difficult, as swings from players with similar skill levels can vary in appearance based on the swing's objective and each golfer's technique. Therefore, we propose using swing videos to predict a more general metric, such as the golf handicap, which reflects a player's overall skill and generally is a good representation of swing quality. To tackle this task, we curate a golf dataset comprising videos of swings accompanied by corresponding golf handicaps. The limited number of diverse players further amplifies the inherent difficulty of this task. Our approach employs a simple architecture, utilizing a CNN backbone to extract

meaningful information from the frames, and a regression head to generate predictions. Given the simplicity of our architecture, our primary investigation revolves around an exhaustive analysis of different problem formulations and losses to address the task. We observe that a relative score regression approach, incorporating a ranking loss, achieves the highest scores in our golf skill assessment problem. Additionally, we explore the utility of incorporating 2D poses as input data to enhance the task's performance.

The remaining sections of the paper are organized as follows: Section 2 provides a comprehensive review of related work on the task of skill determination. Section 3 describes the data collected for the task, and Section 4 explores the different approaches we experimented with. Results are presented in Section 5 and further discussed in Section 6. Finally, Section 7 concludes the paper.

2 RELATED WORK

The domain of video-based skill determination, also known as action quality assessment (AQA), has been relatively underexplored in computer vision research when compared to tasks like action classification or action localization. However, due to its many real-world applications, there has been an increasing interest in this area in recent years, particularly in the sports and medical fields. Several public datasets [21, 20, 22, 10, 28, 18, 3] have been released, focusing on actions within sports disciplines such as diving [21, 20, 22], gymnastics [20, 22, 28], skiing [20], skating [22], surgical activity [10], and rehabilitation exercises [18, 3]. However, there is currently no dataset specifically designed for AQA in the context of golf swings. While there exist works that deal with golf video data, they primarily concentrate on alternative tasks such as golf swing sequencing [19]. Despite the lack of a dedicated dataset for AQA on golf swings, there is a growing need for this kind of evaluation in golf, given the importance of precise and accurate swings for successful performance.

Most video-based AQA methods typically divide videos into clips of predefined lengths and extract features for each clip using a 3D backbone such as I3D [4] or C3D [23]. Moreover, most of the methods focus on a more computationally efficient approach by pre-extracting features using a pre-trained backbone. However, some more recent approaches jointly fine-tune the backbones while training the AQA model, allowing for increased flexibility and learning ability. In our approach, we follow latest approaches and fine-tune the backbone while training the AQA model. In alignment with these recent advancements, our approach also adopts the practice of fine-tuning the backbone while training the AQA model. Additionally, instead of processing the whole video or having to presegment the video into clips, we focus only on the relevant frames, i.e. the clip centered at the time instant when the golf club strikes the ball.

In the realm of AQA methods, we can distinguish three main categories: score regression [7, 20, 25, 22, 26], relative score regression [16, 27, 29], and ranking methods [5, 6, 17].

Score regression. These methods aim to predict a score for each input video. Among these methods, differences primarily lie in the way they aggregate features from different clips before generating regression predictions. For instance, Parmar and Morris [22] compare three different feature aggregation methods: simple average aggregation, an LSTM model with its final output used for aggregation, and an LSTM outputting predicted scores for each step, which are then weighted and averaged. Xu et al. [26] introduce a self-attention mechanism before the LSTM layer to prioritize important clips and reduce the weight of non-important ones. They also propose a multi-scale convolutional skip LSTM to capture sequential information at multiple scales. Additionally, Xu et al. [25] introduce a different scoring paradigm called Likert Scoring, based on the Likert scale. They employ a Transformer encoder-decoder architecture to enrich the clip representation and pass this information to a set of learnable queries in the decoder. Each query is associated with a given score and transformed into a weight between 0 and 1, which is then used to merge the scores associated with each query.

Relative score regression. These methods, instead of directly regressing a score for a given video, take pairs of videos as input and predict the difference in their scores. This approach has proven useful, particularly in small datasets where overfitting is more common. For instance, Li et al. [16] produce both individual regression scores and a relative score for a pair of videos, all of which are evaluated in the loss function. In contrast, Yu et al. [27] only predict the relative score between two videos and further improve performance by introducing a binary tree architecture that classifies the output into different score ranges. A final regression layer refines the relative score within the classified interval. Zhang et al. [29] enhance the previous method by incorporating a module that models the score distribution as a Gaussian distribution instead of predicting a single value for the relative score.

Ranking methods. Ranking methods incorporate a loss function based on correctly ranking a set of given videos. While some methods [5, 6] focus solely on correctly ranking different pairs of examples, Li et al. [17] incorporate the ranking loss into previous regression problems to further improve the correct ordering of predictions.

In our proposed method, we begin by implementing straightforward score regression techniques. Building upon this foundation, we then integrate ranking losses and relative score regression into our model. Extensive evaluation reveals that integrating relative score regression and including ranking terms in the loss function brings benefits for AQA tasks compared to simple score regression approaches.

3 DATA

Our approach to determining golf skill involves utilizing data obtained from participants who have chosen to participate in the Trackman [1] development program. This data includes a video of the golfer executing a golf swing, along with additional information about the player. One of the key pieces of information provided is the golfer's current handicap, which is a measure of their overall skill level in golf. It is essential to highlight that a golfer's handicap is not specifically indicative of their performance in a single swing. Instead, it offers a more general assessment of their consistency

Figure 1: Golf swing progression comparison of two golfers with different handicaps (15 and 5). The top golfer demonstrates consistent balance throughout the swing, while the bottom golfer momentarily loses balance in the last frame. However, when considering the full sequence, the bottom golfer's swing showcases greater power and fluidity, highlighting the importance of analyzing the temporal dynamics in assessing golf swing performance.

and overall proficiency in the sport. However, we believe that by closely examining the characteristics of a golfer's swing captured in a single video, we should be able to gain insights into their overall skill level or at least approximate it. It is important to note that exceptional cases may arise, such as instances where the player fails to make a good impact with the ball or exhibits unusual movements, where the swing may not represent the skill of the player.

The handicap of a player can be defined as an average of the eight lowest score differentials (SD) from their last twenty golf rounds, with lower values indicating a higher level of skill. The score differential can be calculated using Equation 1.

$$\mathrm{SD} = \frac{113}{\mathrm{SR}} \cdot (\mathrm{AGS} - \mathrm{CR} - \mathrm{PCC}_{\mathrm{adj.}}) \tag{1}$$

Here, SR represents the *score rating*, which measures the difficulty of a specific golf course. It has a range of values from 55 to 155, with an average rating of 113. AGS represents the *adjusted gross score*, which is the number of strokes taken to complete a round of golf, adjusted to ensure it never exceeds a net double bogey (i.e., the maximum score for a hole). CR represents the *course rating*, which indicates the expected number of strokes to complete the course. Additionally, $\mathrm{PCC}_{\mathrm{adj.}}$ is an optional adjustment that can be applied if the round is played under unusual conditions, such as extreme weather.

Concretely, the handicap we utilize for our analysis is computed by TrackMan using a golf simulator, which allows players to engage in full rounds of virtual golf. In order to mitigate potential inaccuracies in assessing the player's skill based on the video or perturbations in the handicap, we apply certain filters to the data. Firstly, we exclude players whose estimated handicap is based on fewer than twenty full rounds of golf. This is done to ensure a more reliable estimation of the player's skill level. Additionally, we only consider video clips where the camera is positioned to face the golfer directly, and we focus specifically on swings performed with a driver. The driver is a golf club typically used for the first tee shot, with the objective of maximizing the distance the ball travels. By selecting these specific video clips, we aim to visualize the swing characteristics accurately and maintain a consistent objective across the analyzed videos. This approach facilitates the task by forcing that similar characteristics (e.g., ball impact strength, angle of impact) may lead to similar skill levels. Furthermore, we discard swings with a total carry distance of no more than 100 meters. Considering that professional female golfers typically achieve carry distances between 220 and 255 meters, while the tour average for males is around 275 meters, swings with less than 100 meters indicate inadequate contact with the golf club. Filtering out these outlier strokes helps maintain data integrity. To standardize the data, we mirror the videos of left-handed golfers, ensuring that the progression of the golf swing appears consistent in all videos. We also convert all videos to 30 frames per second and align them so that the frame capturing the moment the golf club strikes the ball remains consistent across all videos. Examples of the resulting processed data can be seen in Figure 1.

Upon data processing, we obtain a dataset comprising 2907 front-facing videos capturing golfers driving the golf ball. Among these videos, there are 284 different subjects, with each subject having a varying number of videos. The median number of videos per player is 8. This relatively small number of unique players introduces an additional challenge to an already complex task, as it increases the risk of overfitting and makes it more difficult to discern the relevant characteristics for skill prediction.

The handicap values within our dataset span from -7 to 28.5, representing a broad range that encompasses both skilled professional golfers and novices. The distribution of these handicap values is visualized in Figure 2. It seems to follow a normal distribution, centered around a mean value of approximately six. However, there is an observable long right tail in the distribution, indicating a relatively high number of golfers with higher handicap values. This right-tail poses an additional challenge when predicting handicaps within this particular range, as the availability of data points

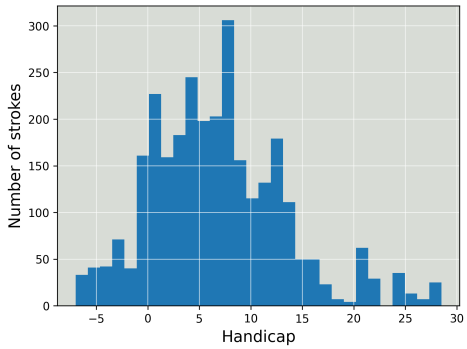


Figure 2: Handicap scores distribution of strokes for all available data.

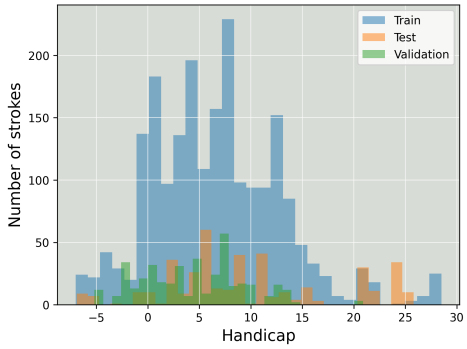


Figure 3: Handicap scores distribution of strokes for the train, validation, and test set.

becomes more limited. Consequently, accurately predicting the handicap values for golfers in this range is more difficult due to the scarcity of data.

In line with standard practice, we randomly divided our dataset into three sets for training and evaluation purposes: a train set (70%), a validation set (15%), and a test set (15%). Importantly, we ensured that all videos featuring a specific subject were grouped together in the same set. The distribution of these sets is illustrated in Figure 3. Upon analyzing the distribution of these sets, we observe that the test split deviates the most from the original distribution. Specifically, it exhibits a higher proportion of high handicap values compared to the other sets. In contrast, the train and validation sets appear to distribute similarly to the original distribution of handicaps. We employ the train set to train our models. The validation set plays a crucial role in monitoring the training process and determining when to stop to prevent overfitting. Finally, the test set serves as the final benchmark for comparing the performance of the different proposed methodologies.

4 METHODS

Our proposed methods rely on two fundamental modules: a 3D backbone and a regression head.

3D backbone. The 3D backbone takes a video input of dimensions $[C \times T \times H \times W]$, where C represents the three color channels, T denotes the temporal dimension, and $H \times W$ indicates the spatial resolution. It generates an embedding of dimension d, that should encode all the relevant information in the video. For the backbone, we employ a straightforward 3D ResNet-18 architecture [8], pretrained on the action classification dataset Kinetics-400 [15], from which discard the classification layer and replace it by the regression head.

Regression Head. The regression head takes the d-dimensional video embedding output by the backbone as input and utilizes a Multi-Layer Perceptron (MLP) with three layers to generate a single score value. This module incorporates dropout with probability of δ, ReLU as intermediate activation functions, and applies a final sigmoid activation function to constrain the output values between 0 and 1. These values represent the predictions for the scaled handicap score y as defined in Equation 2, where HC denotes the original handicap, HCmin represents the minimum handicap, and HCmax denotes the maximum handicap.

$$y = \frac{\mathrm{HC} - \mathrm{HC}_{\mathrm{min}}}{\mathrm{HC}_{\mathrm{max}} - \mathrm{HC}_{\mathrm{min}}} \tag{2}$$

While our architectural approach is simple, our experiments mainly rely on the problem configuration to learn the task and the choice of losses. In the following subsections, we outline our approaches, starting with a fundamental score regression approach. We then expand upon it by incorporating a ranking loss term. Additionally, we explore alternative formulations, such as relative score regression, and investigate the classification of handicaps within different score ranges.

4.1 Score regression

A straightforward approach to solve the task is to just regress the handicap score, which is scaled to a range of 0 to 1 in our case. As mentioned earlier, we can utilize our base models to generate embeddings and then make predictions for each video. To train this model, we employ a loss function defined in Equation 3. In this equation, N represents the number of samples in the minibatch, y_i represents the ground-truth scaled handicap score, $\hat{y}_i$ represents the predicted score, and p is a parameter that allows us to alternate between Mean Squared Error and L1 loss. More specifically, when p is set to 2, we use the Mean Squared Error loss, whereas for p equal to 1, we employ the L1 loss.

$$\mathcal{L}_{\mathrm{Reg.}} = \frac{1}{N} \sum_{i=1}^{N} |y_i - \hat{y}_i|^p \tag{3}$$

4.2 Ranking based loss

In the task of skill determination, it is crucial not only to minimize the prediction error in scoring but also to ensure accurate ranking of predictions. Particularly in applications that involve player comparison, correctly ordering the predictions can be more important than achieving a smaller regression error. To address this, we can further extend our previous approach by introducing an additional

term to our loss function that penalizes incorrectly ordered predictions. As a result, we propose the following loss function, as shown in Equation 4.

$$\mathcal{L} = \beta \cdot \mathcal{L}_{\text{Reg.}} + (1 - \beta) \cdot \mathcal{L}_{\text{Rank.}} \tag{4}$$

Here, $\mathcal{L}_{\text{Rank.}}$ corresponds to the loss component associated with the accurate ranking of different golfers and is defined in Equation 5. By adjusting the β parameter, we can control the emphasis placed on either the ranking or regression aspect of the loss function.

$$\mathcal{L}_{\text{Rank.}} = \frac{1}{\binom{N}{2}} \sum_{i=1}^{N} \sum_{j=i+1}^{N} \max\left(0, -(\hat{y}_j - \hat{y}_i) \cdot \text{sign}(y_j - y_i) + m\right) \tag{5}$$

In the ranking component, we assess all $\binom{n}{2}$ possible pairs of samples and penalize any cases where the ranking is inaccurate. This penalty term takes into account the disparity between predictions. Additionally, we introduce a margin parameter m, which encourages pairs that are correctly ranked but have a small difference in predictions to increase their predicted handicap difference.

4.3 Relative score regression

An alternative, more intricate approach involves redefining the objective of the model in order to predict the handicap difference (i.e., relative score) between each pair of videos within a batch, instead of directly regressing the handicap value of a single video. The structure of a relative score prediction model remains similar to the previous models and can still be trained using the loss defined in Equation 4. As before, we can use a regression term and also a ranking term to ensure that the predicted relative scores are properly ordered. To implement this approach, we follow these steps. First, we extract embeddings from each video in the batch using the same backbone with shared weights. These features are then concatenated for each pair of samples. Next, we pass these concatenated embeddings to the regression head, which predicts the relative score for that particular pair.

During the inference stage, for each example in the test set, we employ a process that involves sampling, n_{samples}, from the training data. These samples serve as reference examples for the test set. The model predicts the relative scores between the test video, denoted as x_{test}, and each of the reference samples. To calculate the handicap score, we add the predicted relative score to the ground truth score of the reference samples. The final prediction for the test example, denoted as $\hat{y}_{\text{test}}$, is obtained by averaging all the predictions across the references, as outlined in Equation 6.

$$\hat{y}_{test} = \frac{1}{n_{\text{samples}}} \sum_{i=1}^{n_{\text{samples}}} y_i + \hat{\Delta}(x_i, x_{test}) \tag{6}$$

Here, n_{samples} represents the total number of reference samples, (x_i, y_i) represents an example sampled from the training data, and $\hat{\Delta}(.,.)$ represents the predicted relative score between the two videos. For a visual representation of the inference step, refer to Figure 4.

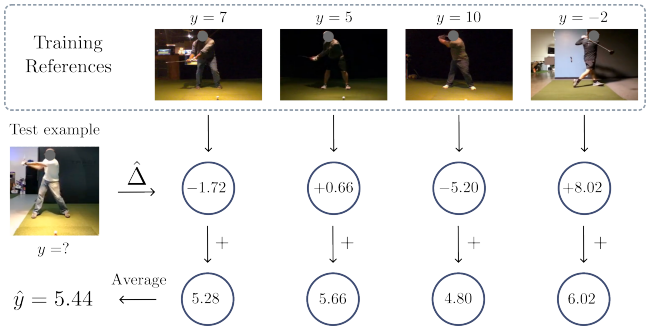


Figure 4: Test time inference for relative score regression. The handicap of a test example is estimated by averaging multiple predictions from reference examples.

4.4 Handicap group classification

An alternative approach to previous methods involves defining the problem as predicting each video into a pre-defined set of handicap groups. By doing so, we transform the task into a more manageable and learnable one for the model. It becomes easier for the model to learn to which interval of handicap values a video can be classified, rather than accurately predicting the exact handicap.

To implement this approach, we first establish the set of handicap groups, $\{[-7; 0], (0; 5], (5; 10], (10; 15], (15; 20], (20; 25], (25; 30]\}$. Then, we modify the regression head in our architecture to incorporate a classification head. The classification head retains a similar structure, with the only distinction being a modification in the last layer. This modification enables the output to have dimensions equal to the number of handicap groups and uses softmax as the activation function. The output provides the probabilities of a video corresponding to each group. Furthermore, the loss function is adjusted to a typical cross-entropy loss, which suits the classification task.

However, this approach encounters two main issues. Firstly, classifying predictions into groups results in a loss of precision in our predictions. Secondly, training the model solely as a classification task disregards the varying degrees of error depending on the distance between the predicted group and the ground truth group. Consequently, a prediction that is off by just one group should be penalized less than a prediction that is further away.

4.5 Ordinal regression

We refer as ordinal regression an approach that builds upon the group handicap classification method to address the issue of neglecting varying degrees of error based on the proximity between predicted and ground truth groups. This approach is inspired by the work of Fuchs and Keshet [9], incorporating a regression-based approach to the classification problem. In this case, our predictions are generated by a regression head, without using sigmoid activation function, to produce handicap score predictions. During training, a special design is implemented to feed each mini-batch with one sample from each handicap group. The loss function evaluates pairs of samples corresponding to consecutive groups. For a pair of samples (x_1, x_2), defined within a set of boundaries (b_1, b_2, b_3) such that x_1 belongs to the handicap interval group $(b_1, b_2]$, x_2 belongs

to $(b_2, b_3]$, and $\hat{y}_1$ and $\hat{y}_2$ represent the predicted scores for each sample, the loss function for each pair of samples is defined as shown in Equation 7. The overall batch loss is calculated as the average across all pairs.

$$\begin{aligned}\mathcal{L}_{(x_1,x_2)} = &\max(0, \gamma + b_1 - \hat{y}_1) + \max(0, \gamma - b_2 + \hat{y}_1) + \\ &\max(0, \gamma + b_2 - \hat{y}_2) + \max(0, \gamma - b_3 + \hat{y}_2)\end{aligned} \tag{7}$$

Here, γ is a parameter that allows adjustment of the interval of correct values (i.e., where the prediction is not penalized) compared to the original group interval. Positive values shrink the interval, while negative values expand it. Predictions outside of this interval are penalized based on their distance from the interval boundaries. Consequently, this loss function enables additional penalties when predictions deviate further from the ground-truth group.

5 RESULTS

In this section, we provide an overview of the implementation and training details for our proposed approaches. Additionally, we present the results obtained from evaluating each approach.

5.1 Implementation and training details

We implemented our model using PyTorch and optimized it with the Adam optimizer, employing a learning rate of $1e^{-4}$. Each video was processed with a total of 40 frames ($T = 40$), each having a spatial resolution of 256×256. The embedding dimension d was set to 256. In models incorporating the ranking term, we set $\beta = 0.95$ and a margin of $m = 0.02$. For relative score regression, we utilized $n_{\text{samples}} = 100$ to generate the predictions. For dropout, we used $\delta = 0.5$.

The models were trained for a maximum of 50 epochs using the train split, and early stopping was performed using the validation split. To evaluate and compare different models, we utilized the test split and two distinct metrics. The first metric employed was the Mean Squared Error (MSE), which quantifies the dissimilarities between the predicted and actual values. The second metric was the percentage of correctly ranked pairs, derived directly from the Kendall correlation and denoted as $\frac{\tau}{2} + 0.5$, where τ represents the Kendall correlation [2, 12]. This metric ensures that our proposed approaches not only generate predictions that closely match the actual values but also maintain the correct ranking of different golfers.

5.2 Experiments

In line with standard regression practices, we begin by establishing a simple baseline for assessing golf skills. This baseline involves predicting all test split golfers' with the mean handicap computed over all videos within our training split. The resulting MSE for this baseline approach is 77.53, which serves as the initial score that we aim to improve upon with our proposed models. Since all predictions are identical under this baseline, calculating the percentage of correctly ranked pairs is not feasible. Next, we evaluate the performance of the different approaches presented in section 4. The results of these evaluations are summarized in Table 1 and a detailed discussion of them can be found in section 6.

Table 1: Performance comparison of models using different problem formulations, evaluated based on mean squared error (MSE) and percentage of correctly ranked pairs. The best performing model for each metric is highlighted in bold.

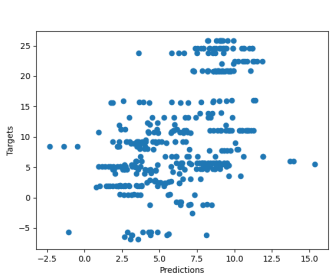
Model	MSE	Correctly ranked pairs (%)
M0: Baseline	77.53	-
M1: Score reg. (L1)	63.88	67.73%
M2: Score reg. (MSE)	67.40	66.70%
M3: Relative score reg. (L1)	59.38	68.12%
M4: Relative score reg. (MSE)	53.49	67.12%
M5: M4 + Ranking	**50.13**	68.00%
M6: Group classification	111.74	61.30%
M7: Ordinal regression	63.12	**68.50%**

6 DISCUSSION

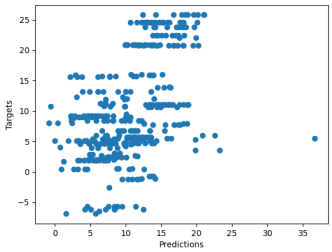
As illustrated in Table 1, all of the proposed approaches, except for the group classification of handicaps, demonstrate improvements over the baseline model in terms of MSE. Notably, the simple approaches of score regression, models M1 and M2, already exhibit an MSE reduction of 10 to 14 points. Among these, the model utilizing the L1 loss (M1) performs better than the MSE loss and achieves a 67.73% accuracy in correctly ranked pairs.

Furthermore, we observe a noticeable enhancement when transitioning from score regression to relative score regression (models M3 and M4). However, in this case, using MSE loss yields better results, reducing the MSE on the test set to 53.49. Nevertheless, there is no noticeable improvement in terms of correctly ranked pairs, as the percentages range between 67% and 68%. Moreover, by incorporating the ranking term into the relative score regression problem in M5, we observe an additional enhancement, resulting in our best MSE score of 50.13. This incorporation also slightly improves the accuracy of correctly ranked pairs to 68%. The superiority of relative score regression over simple score regression can likely be attributed to the following factors:

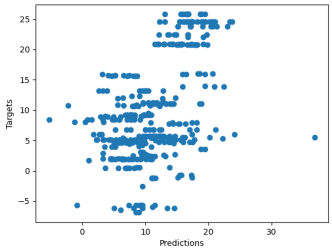
(1) **Increased difficulty of overfitting.** While predicting the handicap of players from a video can be prone to overfitting, predicting the difference between two golfers is a more challenging task. While a model may find it easier to predict the handicap by distinguishing the physical attributes of an individual player rather than their swing technique, this confusion becomes more difficult when considering the influence of both players on the score. In the second case, the model needs to learn more nuanced features and consider the interaction between players, making it less susceptible to overfitting based solely on individual player characteristics. Additionally, the larger number of player pairs to be evaluated compared to individual players helps reduce overfitting.
(2) **More robust test predictions.** During inference, the final handicap prediction is obtained by averaging multiple predictions that use different players as references. This approach enhances the robustness of the solution by mitigating the impact of a few inaccurate predictions or outliers.



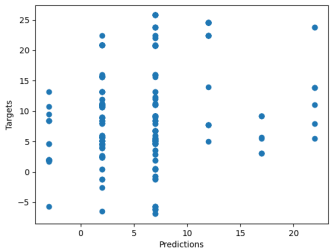
(a) M1 - Score reg. (L1)



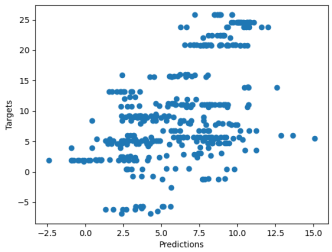
(b) M4 - Relative score reg. (MSE)



(c) M5 - Relative score reg. (MSE) + Ranking



(d) M6 - Group classification



(e) M7 - Ordinal regression

Figure 5: Correlation plots between predictions and ground-truth handicaps for different models presented in Table 1.

In Table 1, we can also observe that the additional experiments involving handicap group approaches yield inferior results in terms of MSE compared to the relative score approach. Specifically, when classifying golfers into different interval groups of handicaps, the MSE is heavily affected as we assign the intermediate point of the interval to all golfers within a group. Furthermore, this approach does not yield satisfactory ranking results. However, introducing ordinal regression yields more favorable outcomes. It achieves a better MSE than simple score regression methods and attains the highest percentage of correctly ranked pairs at 68.50%. In conclusion, our findings highlight the effectiveness of the approach that predicts relative scores while penalizing incorrectly ranked relative scores, resulting in improved outcomes. These results are promising and demonstrate an advancement over the baseline performance. However, it is crucial to acknowledge that the task of golf skill determination still presents ample room for further improvement. In the subsequent subsections, we will delve into a more detailed analysis of our model predictions and analyze the fusion of multiple modalities (raw video and body pose).

6.1 Model predictions

To gain deeper insights into the predictions generated by some of the models presented in Table 1, we conducted a detailed analysis and plotted the predicted values against the ground-truth handicaps in Figure 5. This visualization allows us to explore the correlation between predicted scores and actual handicaps, as well as identify distinct behaviors exhibited by the different methods. Upon visual inspection, it becomes evident that most of the plots demonstrate a positive correlation between the predicted scores and ground-truth handicaps. This suggests that our model effectively captures meaningful information from the video input. An intriguing observation arises when examining the range of predictions across these plots. We noticed that while the correlations remain relatively consistent, there is a substantial variation in the range of the predicted values. Models employing simple regression or ordinal regression approaches (Figure 5a and Figure 5e) exhibit a relatively narrow prediction span, typically reaching up to approximately 15. In contrast, models predicting the relative score (Figure 5b and Figure 5c) display a much broader range of predictions. This discrepancy raises the question of whether models with narrower ranges are truly capturing the full spectrum of skill levels. Conversely, it suggests that relative score approaches possess the capability to predict across a wider range of values, covering nearly all skill levels. These findings further reinforce the superiority of relative score predictions compared to other approaches. They also highlight the importance of developing models that can effectively span the entire spectrum of skill levels. Such models would enable a more comprehensive analysis and facilitate targeted feedback and guidance for golfers, catering to their specific skill levels and needs.

6.2 2D pose modality

In addition to the previously discussed approaches, we also explore the impact of incorporating 2D poses of the players alongside the raw (RGB) videos. This integration aims to assist the model in prioritizing the golfer's movement over other irrelevant information. To extract the 2D joint positions, we utilize a pre-trained HRNet model [24], which provides us with $J = 17$ joint positions for each of the T input frames. Subsequently, we generate a $H/4 \times W/4$ heatmap for each joint and frame, resulting in a pose input of size $J \times T \times H/4 \times W/4$. As depicted in Figure 6, our model treats the pose heatmaps in the same manner as the input videos. Specifically, we employ two distinct backbones, one for each modality. The output

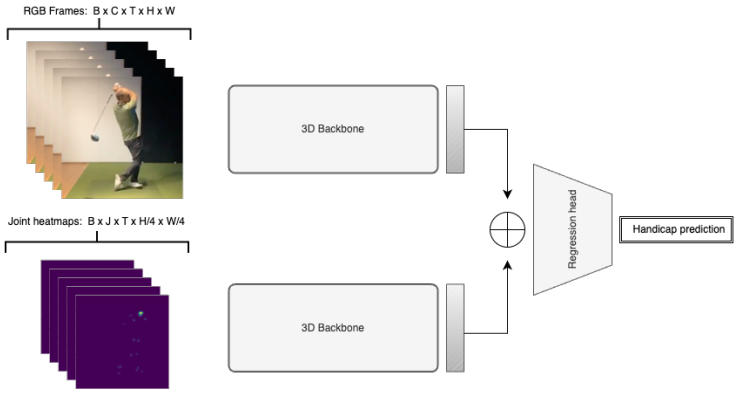


Figure 6: Architecture for fusing multiple modalities (RGB frames and pose heatmaps). Both modalities employ a similar 3D backbone architecture, with separate weights. The output embeddings from each modality are concatenated and fed into the final regression head to predict the skill level of the B golfers in the batch.

embeddings from these backbones are then concatenated before being passed to the final regression head for further processing. Additionally, this experiment is performed using the initial problem definition in subsection 4.1.

Table 2: Comparison of results for the score regression approach (M1) with and without the inclusion of 2D poses.

Model	MSE	Correctly ranked pairs (%)
M1	65.82	67.0%
M1 + Pose	66.10	68.0%

Table 2 presents the results for the models with and without the additional pose input. We observe that when incorporating the pose input, there is a slight decrease in performance based on the MSE, but a slight increase in performance based on the percentage of correctly ranked pairs. Considering these results, we have concluded that including the pose input does not have a clear impact on the overall performance of the model while increasing its complexity. Therefore, we have decided to exclude the pose input from the other models and continue using only the raw videos to predict the player's handicap. We hypothesize that the lack of a clear positive effect when including the poses may be due to the limitations of the predicted poses, as it is a well-known challenge for pose models to accurately capture fast sports movements [13, 14].

7 CONCLUSION

This paper address the challenging task of skill assessment for golfers through swing videos. We have presented a simple model architecture to analyze the performance of various problem design approaches, including simple score regression, relative score regression, the inclusion of ranking losses, and classification-based approaches. Most of these approaches were able to learn relevant information to predict the handicap of a golfer, surpassing the results obtained by a simple baseline that predicts the average handicap for all players. Our observations revealed that an approach predicting relative scores among golfers and incorporating a ranking term yielded the best performance, achieving a MSE of 50. This suggests that, on average, the predictions deviate from the ground-truth values by approximately 7 points. While these results indicate that there is still room for improvement in this task, they also demonstrate that it is indeed possible to extract sufficient information from swing videos to produce a rough estimation of a player's golf skill.

Limitations & Future Work. Despite allowing for a rough estimation of skill, our approach has some limitations that need to be addressed for a more nuanced and accurate assessment. One limitation is the scarcity of different golfers in the dataset. Although we have a reasonably large collection of videos, they only feature 284 unique players, which poses challenges during the training process. Additionally, while we extensively analyzed different design approaches, the simplicity of our chosen architecture may limit its ability to capture finer distinctions in the skill demonstrated in different swing videos. Future work on skill assessment in golf should focus on acquiring a more diverse dataset that includes players of varying skill levels to facilitate the learning process. Furthermore, exploring new architectures specifically tailored to this task, which are more complex and innovative, could help improve performance. Another crucial aspect to consider is the use of explainability techniques to gain a comprehensive understanding of the player's positions and movements, as these factors are essential for achieving a high level of skill. Advancements in these areas could pave the way for developing an AI-driven golf coach that offers personalized feedback and guidance, revolutionizing the accessibility and effectiveness of golf coaching. Our work serves as a foundation for this ambitious goal, providing a basis for future advancements in the field.

8 ACKNOWLEDGEMENTS

This work has been partially supported by the Spanish project PID2022-136436NB-I00 and by ICREA under the ICREA Academia programme.

REFERENCES

[1] TrackMan A/S. [n. d.] Trackman. https://www.trackman.com/. (). Retrieved Feb. 27, 2022 from.

[2] Hervé Abdi. 2007. The kendall rank correlation coefficient. *Encyclopedia of Measurement and Statistics. Sage, Thousand Oaks, CA*, 508–510.

[3] Marianna Capecci, Maria Gabriella Ceravolo, Francesco Ferracuti, Sabrina Iarlori, Andrea Monteriu, Luca Romeo, and Federica Verdini. 2019. The kimore dataset: kinematic assessment of movement and clinical scores for remote monitoring of physical rehabilitation. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 27, 7, 1436–1448.

[4] Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6299–6308.

[5] Hazel Doughty, Dima Damen, and Walterio Mayol-Cuevas. 2018. Who's better? who's best? pairwise deep ranking for skill determination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6057–6066.

[6] Hazel Doughty, Walterio Mayol-Cuevas, and Dima Damen. 2019. The pros and cons: rank-aware temporal attention for skill determination in long videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7862–7871.

[7] Shafkat Farabi, Hasibul Himel, Fakhruddin Gazzali, Md Bakhtiar Hasan, Md Hasanul Kabir, and Moshiur Farazi. 2022. Improving action quality assessment using weighted aggregation. In *Pattern Recognition and Image Analysis: 10th Iberian Conference, IbPRIA 2022, Aveiro, Portugal, May 4–6, 2022, Proceedings*. Springer, 576–587.

[8] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. 2019. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, 6202–6211.

[9] Tzeviya Sylvia Fuchs and Joseph Keshet. 2022. Thor: threshold-based ranking loss for ordinal regression. *arXiv preprint arXiv:2205.04864*.

[10] Yixin Gao et al. 2014. Jhu-isi gesture and skill assessment working set (jigsaws): a surgical activity dataset for human motion modeling. In *MICCAI workshop: M2cai* number 3. Vol. 3.

[11] Malcolm Gladwell. 2008. *Outliers: The story of success*. Little, Brown.

[12] Morten Rieger Hannemose*, Josefine Vilsbøll Sundgaard*, Niels Kvorning Ternov, Rasmus R. Paulsen, and Anders Nymark Christensen. 2022. Was that so hard? estimating human classification difficulty. *Applications of Medical Artificial Intelligence*, 13540, 88. DOI: https://doi.org/10.1007/978-3-031-17721-7_10.

[13] Christian Keilstrup Ingwersen, Janus Nørtoft Jensen, Morten Rieger Hannemose, and Anders B. Dahl. 2023. Evaluating current state of monocular 3d pose models for golf. In *Proceedings of the Northern Lights Deep Learning Workshop*. Vol. 4. DOI: https://doi.org/10.7557/18.6793.

[14] Christian Keilstrup Ingwersen, Christian Mikkelstrup, Janus Nørtoft Jensen, Morten Rieger Hannemose, and Anders Bjorholm Dahl. 2023. Sportspose: a dynamic 3d sports pose dataset. In *Proceedings of the IEEE/CVF International Workshop on Computer Vision in Sports*.

[15] Will Kay et al. 2017. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*.

[16] Mingzhe Li, Hong-Bo Zhang, Qing Lei, Zongwen Fan, Jinghua Liu, and Ji-Xiang Du. 2022. Pairwise contrastive learning network for action quality assessment. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IV*. Springer, 457–473.

[17] Yongjun Li, Xiujuan Chai, and Xilin Chen. 2018. Scoringnet: learning key fragment for action quality assessment with ranking loss in skilled sports. In *Asian Conference on Computer Vision*. Springer, 149–164.

[18] Yalin Liao, Aleksandar Vakanski, and Min Xian. 2020. A deep learning framework for assessing physical rehabilitation exercises. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 28, 2, 468–477.

[19] William McNally, Kanav Vats, Tyler Pinto, Chris Dulhanty, John McPhee, and Alexander Wong. 2019. Golfdb: a video database for golf swing sequencing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 0–0.

[20] Paritosh Parmar and Brendan Morris. 2019. Action quality assessment across multiple actions. In *2019 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 1468–1476.

[21] Paritosh Parmar and Brendan Tran Morris. 2019. What and how well you performed? a multitask learning approach to action quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 304–313.

[22] Paritosh Parmar and Brendan Tran Morris. 2017. Learning to score olympic events. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 20–28.

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

[24] Jingdong Wang et al. 2020. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43, 10, 3349–3364.

[25] Angchi Xu, Ling-An Zeng, and Wei-Shi Zheng. 2022. Likert scoring with grade decoupling for long-term action assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3232–3241.

[26] Chengming Xu, Yanwei Fu, Bing Zhang, Zitian Chen, Yu-Gang Jiang, and Xiangyang Xue. 2019. Learning to score figure skating sport videos. *IEEE transactions on circuits and systems for video technology*, 30, 12, 4578–4590.

[27] Xumin Yu, Yongming Rao, Wenliang Zhao, Jiwen Lu, and Jie Zhou. 2021. Group-aware contrastive regression for action quality assessment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7919–7928.

[28] Ling-An Zeng, Fa-Ting Hong, Wei-Shi Zheng, Qi-Zhi Yu, Wei Zeng, Yao-Wei Wang, and Jian-Huang Lai. 2020. Hybrid dynamic-static context-aware attention network for action assessment in long videos. In *Proceedings of the 28th ACM international conference on multimedia*, 2526–2534.

[29] Boyu Zhang, Jiayuan Chen, Yinfei Xu, Hui Zhang, Xu Yang, and Xin Geng. 2021. Auto-encoding score distribution regression for action quality assessment. *arXiv preprint arXiv:2111.11029*.