Understanding Bouldering using ML Methods

Loïc Comeliau, Julia Heiniger, Weiran Wang
CS433 Machine Learning, EPFL, Switzerland

***Abstract*—Computer vision has been gaining interest in a wide range of research areas in recent years, from medical to industrial robotics [1] and it is one of the most state-of-the-art research topics in Machine Learning nowadays. In this report, we focus on bouldering problems using recorded videos of professional climbers. Using the first 150 frames of a video, we try to make two predictions, i.e. the climber's name and the success of the trial. In our final evaluation, we get 83.9% accuracy for climbers' names prediction and 78.1% accuracy for climbers' success prediction.**

I. Introduction

Competitive bouldering is a new and very popular sport. In 2021, the combined event of climbing was for the first time part of the Summer Olympic Games featuring bouldering together with lead and speed climbing [2]. Nevertheless, there have not been too much on this studies about bouldering yet.

This project belongs to ML4Science [3], an interdisciplinary Machine Learning Project across EPFL campus, which is guided under Swiss Bouldering Olympic Team and Prof. Martin Jaggi. This project focuses on the understanding of bouldering problems from videos, by collecting data, applying pose estimation using computer vision algorithms and then try to make predictions based on our data analysis. Specifically, two predictions were targeted in our project. Our first approach was to use the pose estimation across time in order to predict the climber's name. We then predict whether a climber will succeed to the top by analysing the first 150 frames of climbing videos. The dataset used in this project mainly focuses on 2021 International Federation of Sport Climbing (IFSC) World Cup Series [4].

II. Models and Methods

In order to extract usable data for a Machine Learning model, we first manually labeled some important information, e.g. climbers' name, fail/zone/success which a climber achieve in one video, etc. Then, a pose estimation algorithm is used to get an estimation of the pose of a climber across time. Features are then extracted from these time series evolution of the landmarks. In a further step, the pose estimation will also be normalized relative to the initial pose. For our predictions, we decided to go for the Random Forest classifier, which is a robust and commonly used classifier.

A. Data Processing

The raw data consisted of internal videos of the German national team in World Cups, European Cups and Championships from 2017-2021. Finally, only stable videos, i.e videos that are taken with a static camera position, from the two World Cups in Salt lake city 2021 were considered. In addition, the videos were cut and cropped manually as a preliminary work so that only one climber is visible and only one single attempt is contained in the video. Periods of waiting or cleaning the boulder wall after an attempt were cut off as well. This leads to a dataset of 396 videos. For each one, the following 3 features were manually encoded: the gender of the climber, name of the climber and the achievement. The achievement was split into *fail*, *zone* and *success*, as it is done in boulder competitions. In addition, each video is associated to a boulder problem, which is labeled, to record which videos are from the same boulder problem. In sum, there are 175 videos of female climbers and 221 videos of male climbers, from 23 different climbers with 225 fails, 106 zones and 65 successes.

B. Pose Estimation

The open-source ML solution MediaPipe was used for the pose estimation [5][6]. We also considered OpenPose [7] and OpenPifPaf [8]. In the end, MediaPipe was selected because of its shortest running time, while still providing an accurate pose detection. Note that this choice was made by a subjective analysis by eye. The following parameters were used to apply the mediapipe pose estimation :

- `model_complexity=2` (the highest model complexity is used in order to reach the best pose estimation accuracy)
- `min_detection_confidence=0.5` (default value)
- `min_tracking_confidence=0.5` (default value)

The videos are filmed from different angles and distances to the climbers. Moreover, some body parts are hidden by other body parts, especially small body parts as fingers, feet or the face. Considering this, we decided to estimate the pose of only 15 of the total 32 landmarks suggested by MediaPipe. Here is the list of landmarks we considered most important to track a climber's movement:

- the nose
- upper limbs (left and right): shoulders, elbows and wrists
- lower limbs (left and right): hips, knees, ankles and heels

Applying the pose estimation on each video gives 30 time series (from the 15 landmarks, for each one x- and one y-coordinate). This data will be further processed and then used for future predictions. Note that for some unexplained reasons, there was a problem with MediaPipe on a couple of videos, which leaves in the end 385 processed and ready-to-use videos. In Figure 1 is represented one frame of a video, with the pose landmarks estimated by MediaPipe.

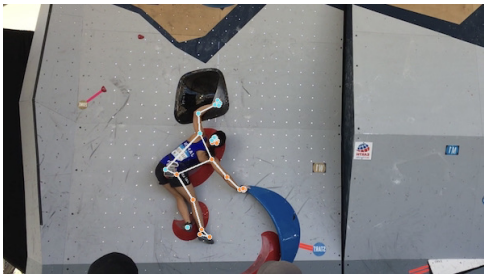

Fig. 1. Example of pose estimation on one frame.

C. Feature Extraction and Selection

To convert time series into useful features, there exists a variety of libraries. We have chosen tsfresh [9], a quite easy to use python solution for time series feature extraction. In a first step, tsfresh calculates several features for each time series, resulting for us in a total number of 23611 features. This is actually quite an enormous number of features, not all of them necessarily relevant for our predictions. To deal with that, tsfresh also provides a method that extracts the most important features for a given label series. Thus, we constructed two subsets of features, one based on the climber's name and the other based on achievement (failure/zone/success).

D. Baseline Predictions

It is interesting to look at some baseline predictions to quantify the improvements done in the next steps.

A first baseline is set by simply using the whole videos duration to predict the climber's name as well as the achievement. The two models are then built using a Random Forest classifier, which gives a first prediction accuracy estimation for the two cases. Here and for the following of the implementations, the Random Forest classifier implementation from the sklearn library has been used [10][11]. Among the different parameters of the random forest classifier, we decided to cross validate the two most important ones, which are listed here:

- `n_estimator = [50, 100, 150, 200, 250, 300, 350, 400, 450, 500]`
- `max_depth = [10, 25, 50, 100, 150]`

Note: we notice in this first tests that the results are strongly dependent on the split of test/train datasets. Therefore, the accuracy of a model is always estimated using 10-fold cross-validation.

A second baseline is computed using only the first frame of the videos for the prediction of the climber's name and achievement. This test is performed to compare it later with the models using only 150 frames. Note that in this case, there is no time series feature extraction and the pose in the first frame is directly used as feature. This model was optimized and tested in the same way as before, i.e. using a 10-fold cross-validation on the Random Forest classifier.

E. Further Predictions

As mention just above, the next step is to make prediction using only the beginning of the video (first few seconds). For example in training sessions of the climbers, this may be useful if it could be applied live to new attempts and one can try to estimate from the beginning of the climbing if the trial will be a success or not. We decided to use the first 150 time points which correspond to around 5-6sec depending on the frame rate of the videos. This period is practically always too short to reach the top and sometimes also too short to reach the zone. We considered thereby only videos with at least 150 frames. Leading us to a sample size of 329 videos. Note that to follow what is done for the prediction of the achievement, the prediction of the climber's name will also be performed using the 150 first frames in the following tests.

At this stage, it is quite interesting to have a simple linear model that would be used as a reference to evaluate the improvement provided by the Random Forest classifier. This reference will then be computed using a Linear Support Vector Classification (also provided by the sklearn library [12]).

As described in Section II-C, we extracted once the filtered features based on the climber's name and once the filtered features based on the achievement using tsfresh. A Random Forest classifier with 10-fold cross-validation of the two specified parameters was then applied again, as well as the aforementioned Linear Support Vector classifier.

F. Data Augmentation

In order to further enhance the predictions and to avoid over-fitting, one requires more samples, i.e. more videos. The idea is therefore to perform data augmentation and there exists several possible methods for time series data augmentation. Considering the work of Iwana and Uchida [13], the Discriminative Guided Warping (DGW) method, provided by them, was thought to be corresponding to what we needed. Actually, this consists of time warping and pattern mixing. Thereby it is guided by a discriminative teacher time series that depends on the labels [14]. The data augmentation was therefore applied to the 150 Frames time series. The method provided by the publication cited here above allows us to get a second dataset of the same size of the initial one. This new dataset is then append to the initial dataset, which gives a augmented dataset which has now a doubled size compared to the previous one. The augmented dataset is then used to build a model in the same way as it has been done before (i.e. feature extraction, Random Forest classifier with 10-fold cross validation).

G. Pose normalization

The comparison between the 1 Frame and 150 Frames model has highlighted that the predictions depends too strongly on the start position, i.e. the 150 Frame model does not provide a significant improvement (see Table I). To get around this problem, we normalize the time series relative to the first frame. Therefore, the coordinates of each landmark in the first frame are subtracted to the coordinates of the other frames in the time series. This way, the start positions of each landmark is (0,0) in the first frame. Again using the same method as in Section II-C, we extracted features depending on the climber's name and on the achievement from the first

150 frames of the normalized time series. With the same 10-fold CV as before we evaluate also the Random Forest lassifier for those two models.

For our final model, we used the normalized data that has been also augmented using DGW, as explained in subsection II-F). This will set the final model we reached by following the different thinking steps explained in this section.

III. Results

The different steps described in the Methods section (II) have been tested in practice and the accuracy have been quantitatively evaluated in each case. This quantitative information is then used to answer to two main questions, for each prediction model (i.e. climber's name and achievement) :

- How is the accuracy of the model changing across the models improvements;
- How accurate is the final model used for prediction.

Recall that the models were evaluated and optimized using 10-fold cross-validation on two parameters of the Random Forest classifier and on the test/train dataset split. In Table I one can find the obtained mean test accuracies over the 10 dataset splits for the different models which are built using the Random Forest classifier. An important note is that for each model, the mean training accuracy is equal to 1.0, except in the case of the 150 Frame normalized climber's name prediction and 1 Frame achievement prediction, where the training accuracy are respectively equal to 0.174 and 0.986.
As mentioned in subsection II-E, a linear model for the non augmented and non normalized dataset has been performed as a reference for comparison. The results of this model are quite bad as the test accuracy for the climber's is around 0.16 (train accuracy of 0.23) while the test accuracy for the achievement is about 0.45 (train accuracy of 0.51).

IV. Discussion

Now that we have a quantitative information about the evolution of the accuracy across the different models, it is interesting to interpret this evolution. Moreover, some investigation on the final model will be performed in order to try to understand the results of this model and evaluate its actual validity.

A. Models improvements

The aim of the project was to use the beginning of the video (i.e. first 5s) to make prediction. The two baseline predictions that have been computed allow us to evaluate the results of the first prediction using the 150 frames (without data augmentation and normalization).

By using the full time series, the reached accuracy for predicting the achievement was already really high (80%). However, the accuracy for the name prediction was very low (around 34%). Nevertheless, the training accuracy is equal to 1.0, which indicates that the model is overfitting the training data. For the second baseline of prediction (using 1 frame), we get an accuracy of 47% for the name prediction, which is actually better than using the full time series. On the other hand, the accuracy of 57% for the achievement prediction is worse than before.

It is now interesting to compare the results of these two baselines with the result of the model using only the first few seconds of the video. As expected, the accuracy for predicting the achievement is worse than using the full time series. However it is not considerably higher as using only one frame. This comparison allows us to assume that when we do not use the whole video, the prediction depends strongly on the starting position of the climber, i.e. the pose estimation in the first frame of the video. This observation could be one of the cause of the overfitting that we have for now.

To handle this problem of overfitting, two approaches were applied. The first idea was to increase the number of samples, as 329 is not so high and this small amount of datapoints could be a cause of ovefitting. Since the preparation of additional data (i.e. video processing by hand) needs too much time for this project, the 150 Frames time series were augmented by a method using DGW (see section II-F). This model reached a clearly higher accuracy for the achievement (80%) and the climber's name (88%). After all, the model still has a training accuracy of 100% for both predictions and the overfitting is not solved.

The second idea to solve overfitting was to standardize the time series with respect to the start position (see II-G) to avoid the dependency on the first frame. The accuracy of predicting the name was drastically decreased thereby (15%). For the first time the training accuracy was not 100%, but unfortunately it was now very low (17%). So that this model is not useful at all. However, the achievement prediction with the normalized time series is still as accurate as with the original 150 Frames (58%). The disadvantage is that it still suffers from overfitting (100% training accuracy). In a last step, the normalized data was augmented. Like that, the final result have been improved to 84% for the climber's name and to 78% for the achievement. Although the training accuracy is still 100% this model is most likely more accurate and robust, which also increases the reliability to the prediction values.

B. Final model evaluation

To further understand the final model, some analyses were performed on it to investigate the quality of prediction. If we simply look at the numbers, the prediction accuracies around 80% seem quite good. However, a big problem in our data set is that it is very unbalanced. Indeed, there are a lot of fails, less zones and only a few successes. Moreover, some climbers are also over-represented and only around 45 different boulder problems are analysed.

First, a summed confusion matrix was created over all 10 folds, showing which names were correctly or wrongly predicted (see appendix, Figure 4). It can be seen that when a name was predicted incorrectly, it was most often assigned to Alexander Megos. This is the climber with the most videos. Other climbers who are represented in many videos were also predicted too often. Another observation we could make is that the wrongly predicted names do not match the gender.

Model	Climber accuracy	n_estimator	max_depth	Achievement accuracy	n_estimator	max_depth
Baseline Predictions						
Full Time series	0.343	500	25	0.803	500	10
1 Frame	0.499	200	25	0.567	150	10
Further Predictions						
150 Frame	0.44	150	25	0.583	500	25
Improved Predictions						
150 Frames augmented	0.881	500	50	0.804	450	25
150 Frames normalized	0.146	100	10	0.58	350	10
150 Frames augmented & normalized	0.839	400	25	0.781	150	25

TABLE I
MEAN TEST ACCURACY FOR CLIMBER'S NAME AND ACHIEVEMENT PREDICTION

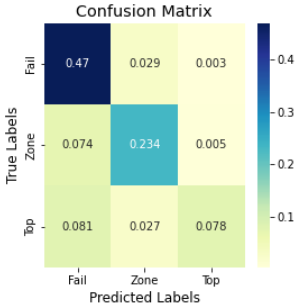

Fig. 2. Confusion matrix of achievement prediction.

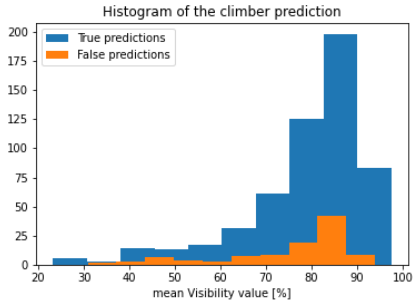

Fig. 3. Distribution of visibilities in correct and wrong climber's name predictions.

This matrix has been computed for the achievement prediction as well. As could be expected, also here, fails have been to too much predicted. This can be observed in Figure 2, where one can see that fail and zone are quite well predicted while the successes are more wrongly than correctly predicted.

Mediapipe allows us to record, for each video analysed, a mean visibility (between 0 and 1) of each landmark across the whole video. We are then able to have an estimation on the pose estimation quality by computing the mean of all landmarks visibility for each video. An idea of statistical analysis of the predictions is to look at the distribution of the mean visibilities for the wrong and the correct predictions, to see if maybe we could have a correlation between wrong predictions and small visibility. This analysis has been performed for both the climber's name and achievement prediction. In Figure 3 are represented histograms of the mean visibility for the correct and the wrong predictions, for the climber prediction. There is not a big difference between the two distributions. However, we can still observe that the ratio of low visibility videos relatively to the high visibility videos seems higher in the wrong predictions. Thereby, this observation could be one of the factors that cause the wrong predictions but it is surely not the primary reason.

Note that on the histograms we can see some videos we really small visibility values (0.2 - 0.5). Unfortunately, it is when doing this final analysis that we made this observation and we did not have time to run again the whole feature extraction and model building with the videos with a mean visibility higher than some threshold. However, this could be tested in the future and it could be interesting to see if the results are sensitive to this change. This could be good as the models are overfitting, these poses may induce wrong predictions. However, this would reduce the size of our dataset which is not so large.

V. SUMMARY

After extracting the time series of almost 400 pre-processed stable boulder videos, the goal was to predict the climber's name and the achievement based on those. At the end, the aim was to have an accurate prediction only from the first few seconds of the videos. The best model was obtained by augmenting the normalized position and then built using a Random Forest Classifier with the parameters `n_estimator = 150` and `max_depth = 25`, reaching an accuracy of 78% to predict the achievement. ..

What makes the task really hard, is that a boulder problem consists of different parts. A climber can easily accomplish the first one, but have more problem with the next one, they do not have to depend on each other. In addition, each climber has strengths and weaknesses, so that some obstacles are simple for one climber but hard for the other one. So, it's difficult to get a robust prediction within 150 frames. However, it would be very powerful and useful in training. It would be really interesting to improve the prediction and further investigate this problematic. In a future project, it would be useful to extend the dataset. So that it is more balanced, i.e. the number of fails, zones and tops is about the same, there are more different climbers represented and to have more different boulder problems.

REFERENCES

[1] A. Voulodimos, N. Doulamis, A. Doulamis, and E. Protopapadakis, "Deep learning for computer vision: A brief review," *Computational intelligence and neuroscience*, vol. 2018, 2018.

[2] C. Lutter, T. Tischer, and V. R. Schöffl, "Olympic competition climbing: the beginning of a new era—a narrative review," *British Journal of Sports Medicine*, vol. 55, no. 15, pp. 857–864, 2021.

[3] M. L. CS-433, "Ml4science," https://www.epfl.ch/labs/mlo/ml4science/.

[4] ifsc climbing.org, "2021 ifsc world cup series," https://www.ifsc-climbing.org/index.php/component/tags/tag/2021-ifsc-world-cup-series.

[5] "MediaPipe." [Online]. Available: https://mediapipe.dev/

[6] V. Bazarevsky and I. Grishchenko, "On-device, real-time body pose tracking with mediapipe blazepose," *Google AI Blog*, 2020.

[7] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "Openpose: realtime multi-person 2d pose estimation using part affinity fields," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 1, pp. 172–186, 2019.

[8] S. Kreiss, L. Bertoni, and A. Alahi, "Openpifpaf: Composite fields for semantic keypoint detection and spatio-temporal association," *arXiv preprint arXiv:2103.02440*, 2021.

[9] "tsfresh — tsfresh 0.18.1. documentation." [Online]. Available: https://tsfresh.readthedocs.io/en/latest/#

[10] "sklearn documentation - randomforestclassifier." [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

[11] Breiman, "Random forests," *Machine Learning, 45(1), 5-32*, 2001.

[12] "sklearn documentation - linearsvc." [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html?highlight=linear%20svc#sklearn.svm.LinearSVC

[13] B. K. Iwana and S. Uchida, "Time series data augmentation for neural networks by time warping with a discriminative teacher." [Online]. Available: http://arxiv.org/abs/2004.08780

[14] "Time series augmentation," original-date: 2020-04-14T12:19:18Z. [Online]. Available: https://github.com/uchidalab/time_series_augmentation

APPENDIX

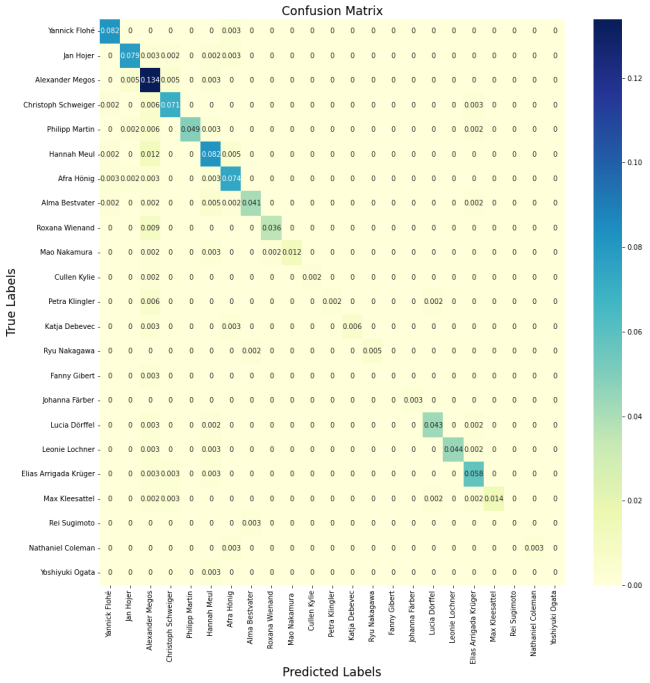

Fig. 4. Confusion matrix of climber's name prediction.