Calisthenics Skills Temporal Video Segmentation

Antonio Finocchiaro, Giovanni Maria Farinella[a] and Antonino Furnari[b]
Department of Mathematics and Computer Science, University of Catania, Italy

Keywords: Pose Recognition, Sports Video Analysis, Temporal Video Segmentation.

Abstract: Calisthenics is a fast-growing bodyweight discipline that consists of different categories, one of which is focused on skills. Skills in calisthenics encompass both static and dynamic elements performed by athletes. The evaluation of static skills is based on their difficulty level and the duration of the hold. Automated tools able to recognize isometric skills from a video by segmenting them to estimate their duration would be desirable to assist athletes in their training and judges during competitions. Although the video understanding literature on action recognition through body pose analysis is rich, no previous work has specifically addressed the problem of calisthenics skill temporal video segmentation. This study aims to provide an initial step towards the implementation of automated tools within the field of Calisthenics. To advance knowledge in this context, we propose a dataset of video footage of static calisthenics skills performed by athletes. Each video is annotated with a temporal segmentation which determines the extent of each skill. We hence report the results of a baseline approach to address the problem of skill temporal segmentation on the proposed dataset. The results highlight the feasibility of the proposed problem, while there is still room for improvement.

1 INTRODUCTION

The discipline of calisthenics is composed of various categories, including Skills, Endurance and Streetlifting[1]. The skills category is the most popular and appreciated among those. It focuses on mastering challenging poses and movements that require high levels of strength, tendon stability, balance and coordination, engaging multiple upper and lower muscle groups simultaneously. In calisthenics, evaluation is generally performed by estimating the duration of a specific skill's hold. Tools able to automatically segment a video in order to identify the execution of a skill and quantify its duration may be useful to support athletes in their training and judges during competitions.

Although previous works have mainly focused on a range of team sports such as soccer (Giancola et al., 2022), basketball (Khobdeh et al., 2023), tennis (Vinyes Mora and Knottenbelt, 2017) as well as individual sports such as swimming (Giulietti et al., 2023), badminton (Rahmad et al., 2020) or yoga (Suryawanshi et al., 2023) in this work, we consider the problem of calisthenics skills temporal video action segmentation (see Figure 1), which can be used to provide assistance to calisthenics athletes during training or judges during competitions. This task has been accomplished by analyzing the 2D body pose of the athletes. In addition, previous works have not explicitly investigated algorithms for temporal segmentation of skills from video, which would be the core of such automated tools. Aiming to provide an initial investigation on this topic, in this work, we contribute with a labeled dataset of videos of athletes performing calisthenics skills. Specifically, we selected 9 skills based on their popularity among amateurs and professional athletes. The dataset contains 839 videos of athletes performing skills, which have been collected from different sources including social networks and ad-hoc recordings, in order to ensure a realistic and natural set of video examples.

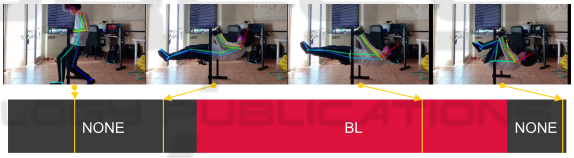

Figure 1: Calisthenics skill temporal video segmentation consists in breaking down a video into segments to highlight the beginning and end of each performed skill. As shown in the figure, the pose of the athlete has an important role in the considered task.

[a] https://orcid.org/0000-0002-6034-0432
[b] https://orcid.org/0000-0001-6911-0302
[1] https://en.wikipedia.org/wiki/Calisthenics

Finocchiaro, A., Farinella, G. and Furnari, A.
Calisthenics Skills Temporal Video Segmentation.
DOI: 10.5220/0012400600003660
Paper published under CC license (CC BY-NC-ND 4.0)
In *Proceedings of the 19th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2024) - Volume 2: VISAPP*, pages 182-190
ISBN: 978-989-758-679-8; ISSN: 2184-4321
Proceedings Copyright © 2024 by SCITEPRESS – Science and Technology Publications, Lda.

All videos have been manually labeled with temporal segmentation annotations which indicate the beginning and end of each skill execution. We hence extracted the spatial coordinates of the athletes' joints in the videos using OpenPose (Cao et al., 2019). To provide initial results on the proposed dataset, we design a simple pipeline which includes a module performing a per-frame classification of body joint coordinates, followed by a module performing temporal reasoning on top of per-frame predictions. The results highlight the feasibility of the proposed problem, while there is still room for improvement. The dataset and code related to this paper are publicly available at the following URL: https://github.com/fpv-iplab/calisthenics- skills-segmentation.

2 RELATED WORKS

This research is related to previous investigations in the fields of computer vision for sports analysis, human action recognition, and temporal video segmentation.

2.1 Computer Vision for Sports Analysis

In recent years, the application of computer vision based techniques for sports analysis has played a fundamental role in the development of automated tools capable of analyzing matches, providing statistics, or assisting the referees during competitions. Most of the previous works focused on team sports such as Soccer (Spagnolo et al., 2014; Manafifard et al., 2017; Banoth and Hashmi, 2022; Huang et al., 2022; Garnier and Gregoir, 2021), Basketball (Zandycke et al., 2022; Xiao et al., 2023; Hauri et al., 2021; Ahmadalinezhad and Makrehchi, 2020; Yoon et al., 2019; Ramanathan et al., 2016), Hockey (Koshkina et al., 2021) and many others (Martin et al., 2021), (Pidaparthy et al., 2021), or individual disciplines such as Diving (Murthy et al., 2023), Table Tennis (Kulkarni and Shenoy, 2021) or Darts (McNally et al., 2021). These works considered different image or video understanding tasks, including detecting and tracking objects and athletes (Liu et al., 2021; Rahimi et al., 2021). The reader is referred to (Naik et al., 2022) for a review of video analysis in different sports. Despite these advances, previous works did not consider the calisthenics field, hence resulting in a lack of datasets, tasks definitions and approaches. In this work, we aim to contribute an initial dataset and a baseline for the segmentation of calisthenics skills from video.

2.2 Human Action Recognition

Human Action Recognition (HAR) is a field of Computer Vision aiming to classify the actions performed by humans in a video. It consists of two main categories as discussed in (Yue et al., 2022; Ren et al., 2020):

- Skeleton-based recognition consists in studying the spatio-temporal correlations among various patterns of body joints. The spatial information provided by the joints can be extracted by a human pose estimation algorithm (Munea et al., 2020; Wang and Yan, 2023). One possible implementation is related to graph convolutional networks as discussed in (Fanuel et al., 2021).
- RGB-based approaches use a different method to detect people, based on the analysis of RGB data within the images (Shaikh and Chai, 2021). Although this approach can be trained end-to-end from videos, it generally needs to deal with the processing of irrelevant information (e.g., the background).

We observe that, in calisthenics skills video temporal segmentation, the athlete body pose plays an important role, while the background is less relevant. Hence, we base our analysis on the body joints extraction using OpenPose (Cao et al., 2019).

2.3 Temporal Video Segmentation

Temporal Video Segmentation is an essential task in the field of video understanding (Richard and Gall, 2016; Cheng et al., 2014; Zhou et al., 2008). It consists in dividing a video into relevant segments that represent the occurrence of predetermined events, such as human actions. The main task is to identify the boundaries of the events in the video in terms of timestamps or frames. In video action analysis, the application of this method allows us to analyze the actions performed by a human subject and their evolution over time. An introduction to this topic, the main adopted techniques and the most used evaluation metrics are discussed in (Ding et al., 2023). In this work, we consider the temporal video segmentation problem of segmenting calisthenics skill execution from video as a mean to estimate the starting, ending, and duration time of each skill. We base our experiments on the approach presented in (Furnari et al., 2018), which factorizes temporal video segmentation into per-frame processing and probabilistic temporal reasoning on top of per-frame predictions. We further compare this approach with a method based on heuristics for the reconstruction of the temporal segmentation of skills.

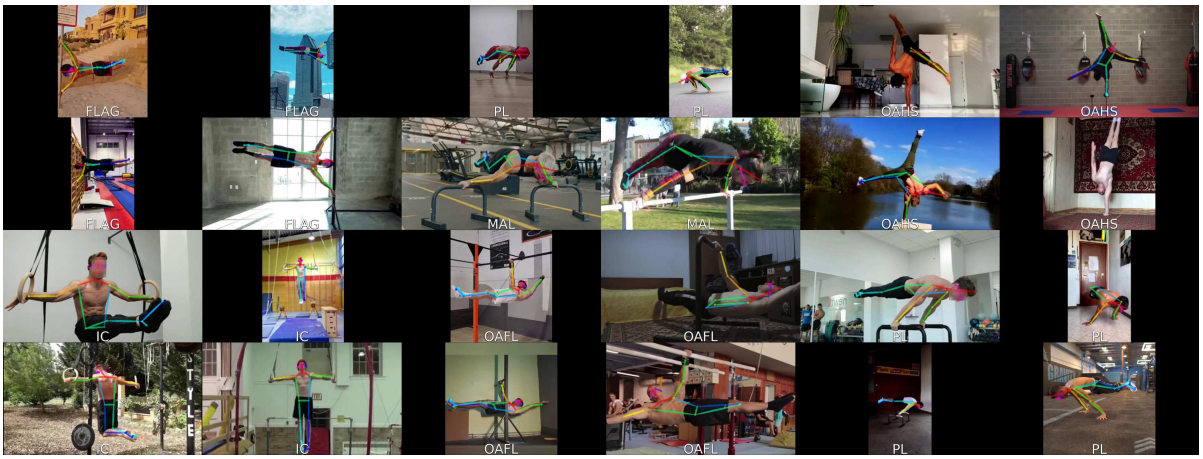

Figure 2: Example frames and related body poses from the proposed dataset.

3 DATASET

In order to create a video dataset, we selected a set of skills to be covered. Each skill has a different difficulty level and demands a specific strength to be performed. All the chosen skills have been chosen based on their popularity among amateur and professional athletes and their importance in competitions such as the following: Burningate[2], SWUB[3], WOB[4], BOTB[5], WSWCF[6]. A detailed description of each skill is given in (Low, 2016). The selected skills are listed in Table 1 together with the collection statistics. We collected a total of 839 videos covering the 9 selected skills. Each video has been converted to a resolution of 960x540 pixels at 24 frames per second and trimmed so that each video comprises one skill. Each video can contain some 'NONE' segments before or after the skill execution (see Figure 1) which models crucially have to recognize to quantify the actual duration of a skill. Almost all videos contain only the athlete in the scene, with few exceptions where multiple people are present in the background. In such cases, the athlete will still cover the foreground area. The most common backgrounds in the scenes include outdoor areas such as parks or streets, as well as in indoor locations like gyms or the athletes' homes. Figure 2 shows some example frames from the proposed dataset. The average duration of skills in the videos is 5.83 seconds, the longest video has a duration of 27.83 seconds, whereas the shortest video lasts 0.83 seconds. The duration of the videos is strongly related to the complexity of the skill performed and the level of the athlete. For each collected video, we have manually labeled start/end times of the skill in frames and seconds, the corresponding skill category label, the MD5 checksum of the file and a video identifier composed of the skill name and a progressive number.

Table 1: Occurrences of each skill in our dataset in terms of number of videos, seconds and frames.

Skill	Videos	Seconds	Frames
Back Lever (BL)	88	574.66	13792
Front Lever (FL)	108	443.08	10634
Human Flag (FLAG)	75	633.16	15196
Iron Cross (IC)	77	513.08	12314
Maltese (MAL)	98	392.70	9425
One Arm Front Lever (OAFL)	80	363.50	8724
One Arm Handstand (OAHS)	94	606.45	14555
Planche (PL)	103	485.25	11646
V-sit (VSIT)	116	814.87	19557
Total	839	4826.79	115843

To allow research on calisthenics skill recognition and temporal segmentation through body pose analysis, we extracted body poses from each frame through OpenPose (Cao et al., 2019). We set the 'number_people_max' flag to 1 to allow the model to detect only the most prominent human when multiple subjects are present in the scene. This does not ensure that the recognized person is the athlete in a multi-person scene, so the video should contain only the athlete for optimal system performance. The 'net_resolution' parameter is set to 208.

We considered the BODY_25B model which can identify up to 25 human joints and provides three numerical values for each joint: the X coordinate, the Y coordinate and the confidence level. Body joint coordinates have been normalized by the frame dimensions in order to obtain values independent of the

[2]https://www.burningate.com/gare-calisthenics
[3]https://streetworkoutultimatebattles.com
[4]https://worldofbarheroes.com
[5]https://worldcalisthenics.org/battle-of-the-bars
[6]https://wswcf.org/competitions

video resolution ranging from 0 to 1. As a result, each frame is associated to 75 numerical features comprising X and Y coordinates of all body joints and related confidence scores.

We divided the dataset randomly assigning 80% of the videos to the training split and 20% of the videos to the test split. The dataset and pre-extracted joints is available at https://github.com/fpv-iplab/calisthenics-skills- segmentation.

4 METHOD

In this section, we describe the baseline calisthenics skills temporal segmentation approach used in our experiments. The proposed method is composed of two main modules: 1) a frame-based multiclass classifier which takes as input body joint features and predicts whether the current frame contains one of the skills or a 'NONE' background segment; 2) a temporal segmentation module which refines the per-frame predictions in order to obtain coherent temporal segments.

4.1 Multiclass Classifier

This component is implemented as a Multilayer Perceptron. Its architecture consists of the first layer, which comprises 75 neurons, followed by three hidden layers, each composed of a linear layer and an activation layer. The chosen activation function is the LeakyReLU, which is shown to outperform other popular activation functions in the experiments. The output layer has 10 nodes, corresponding to the 9 skill classes, plus an additional 'NONE' background class. We train this module with a standard cross-entropy loss and Adam optimizer (Kingma and Ba, 2017). We use a batch size of 512 and train the model over 500 epochs, with the optimizer learning rate set to 0.0001. Optimal hyperparameter values were determined by cross-validation.

4.2 Temporal Segmentation Module

The temporal segmentation module works on top of the predictions of the multiclass classifier to produce coherent temporal segments. The goal is to discriminate skill patterns and reconstruct the video timeline, trying to correct any mistaken skill predictions from the classifier. We consider two versions of this module: one based on a heuristic method and another one based on a probabilistic approach.

4.2.1 Heuristic-Based Temporal Segmentation

The heuristic approach aims to obtain coherent temporal segments from frame-wise predictions in three steps.

Sliding Window Mode Extractor (SWME). This step relies on a sliding window process that iteratively returns the mode of the group of frames within the window.

Given the sequence of all n frames present in the video, $F = [f_0, f_1, \ldots, f_k, \ldots f_{n-1}]$, the base window size is defined as:

$$w_b = \lfloor ((1-s) \cdot m \rfloor$$

with $m = 32$ and s defined as follows:

$$s = 0.5 + \sum_{i=0}^{n-2} (up[F_i = F_{i+1}] + dw[F_i \neq F_{i+1}])$$

where $up = \frac{0.14}{n}$, $dw = -\frac{0.11}{n}$. The best-performing values for these constants have been defined following a naive approach of trial and error where up and dw represents respectively a reward and a penalty factor. The m value consists of a multiplier factor which adjusts the window range from $[0.39, 0.64]$ to $[11, 19]$. Through these assignments, the size of the window is set as an inversely proportional ratio to the frame variance of the video. As can be seen, the divergences between contiguous frames are less weighted than equally labeled frames. This is related to the high frequency of different sequences of frame classes. We hence process the video with a sliding window of size w_s which is initially set to $w_s = w_b$. Formally, the sliding window approach identifies subsets of F:

$$V^{(k)} = [f_{c-w_s+1}, f_{c-w_s+2}, \ldots, f_c] \ \forall c \in \{0, n-1\}$$

The apex represents the k^{th} subset iteratively taken in F. Thus, we compute the mode of the considered subset. If more than one mode is found in $V^{(k)}$, there is no agreement in the current window, and we enlarge it by incrementing w_s and c by one unit (assuming that they do not exceed n). The mode-seeking process is hence iterated. When a single mode is found, w_s is reset to $w_s = w_b$ and c is incremented by $w_s - stride$ (we set $stride = 3$) to enlarge the window by one unit and calculate again the mode. For each step, we store the following three attributes about the local mode:

$$idx_{start}(f_i)^{(k)} = min\{j | f_j^{(k)} = f_i^{(k)}\}$$

$$idx_{end}(f_i)^{(k)} = max\{j | f_j^{(k)} = f_i^{(k)}\}$$

$$mode^{(k)} = f_i^{(k)} \in Mode^{(k)} \iff |Mode^{(k)}| = 1$$

These operations are repeated for each element in F,

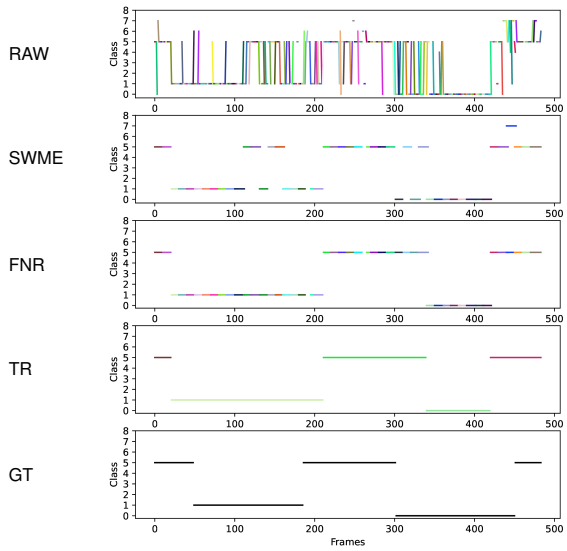

Figure 3: A representation of the heuristic algorithm applied to MLP predictions on a sample video. In the prediction of our algorithm, different colors represent different segments.

eventually resulting in the set:

$$W = \{I_0, I_1, I_2, \ldots, I_k, \ldots, I_{w-1}\}$$
$$\text{where } w = |W| \text{ and}$$
$$I_k = \{mode^{(k)}, idx_{start}(mode^{(k)})^{(k)}, idx_{end}(mode^{(k)})^{(k)}\}$$
$$\forall k \in F, h \in W$$

It is worth noting that we obtain $k \ll n$, where $n = |F|$. The output of this step consists of a smaller list of elements, each of which contains the three defined attributes. These are initial candidate video segments. Figure 3 compares the results obtained in this step (second row) with the raw predictions (first row). As can be noted, this step significantly reduces the noise in the first row list, but may still contain incorrect segments due to high uncertainty.

Filtering and Noise Removal (FNR). This step focuses on decreasing noise, applying a slightly modified version of the previous step. Given the set W from the previous step, this second stage returns a modified set, denoted as R:

$$R = \{I'_0, I'_1, \ldots, I'_j, \ldots, I'_{w-1}\}$$

Where the j^{th} element is transformed as follows:

$$I'_j = \text{mode}(I_{\max(j-2,0)}, I_{\max(j-1,0)}, I_j, I_{\min(j+1,w-1)}, I_{\min(j+2,w-1)})$$

As observed in the third row in Figure 3, some incorrect segments belonging to class 1, are effectively replaced with the local segment mode, providing a reliable reconstruction of the skill segment.

Timeline Reconstructor (TR). This step defines a new set denoted as T:

$$T = \{I''_0, I''_1, \ldots, I''_k, \ldots, I''_{t-1}\}$$
$$\text{with } t < w$$

Where the I''_k element is defined as follows:

$$I''_k = \begin{cases} I'_j, & \text{if } x_j \neq x_z \, with \, x_j \in I'_j, x_z \in I'_{j+1} \\ merge(I'_j, I'_p), & \text{otherwise} \end{cases}$$

With the *merge* function defined as follows:

$$merge(I'_j, I'_p) = \{x_j, idx_{start}(x_j), idx_{end}(x_p)\}$$
$$\forall j < p, j, p \in R$$

Through this step, different segments belonging to the same class are combined into a single segment representing a skill. The effect is illustrated in the fourth row in Figure 3.

4.2.2 Probabilistic-Based Temporal Segmentation

The probabilistic temporal segmentation module aims to output coherent temporal segments from per-frame predictions assuming a simple probabilistic model in which the probability of two consecutive frames having different classes is assumed to be low. We follow an approach similar to (Furnari et al., 2018). Let $F = \{f_1, \ldots, f_N\}$ be an input video with N frames f_i and let the corresponding set of labels be denoted as $L = \{y_1, \ldots, y_N\}$. The probability of labels L is modeled assuming a Markovian model:

$$P(L|F) \propto \prod_{i=2}^{n} P(y_i|y_{i-1}) \prod_{i=1}^{n} P(y_i|f_i).$$

Where the term $P(y_i|y_{i-1})$ represents the probability of transiting from a per-frame class to another. The transition probability is defined as follows:

$$P(y_i|y_{i-1}) = \begin{cases} \varepsilon, & \text{if } y_i \neq y_{i-1} \\ 1 - M\varepsilon, & \text{otherwise} \end{cases}$$

The constant ε is hence defined as: $\varepsilon \leq \frac{1}{M+1}$. Finally, the global set of optimal labels L is obtained maximizing $P(L|F)$ using Viterbi algorithm:

$$L = \arg\max_{L} P(L|F).$$

5 EXPERIMENTAL SETTINGS AND RESULTS

In this section we report experimental settings and results on the main components of the proposed pipeline.

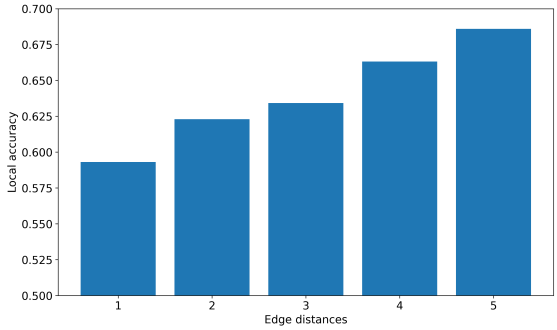

Figure 4: Accuracy (y axis) versus distance of the classified frame from the edges of ground truth segments (x axis).

5.1 Multiclass Classifier

The optimal configuration for the Multilayer Perceptron (MLP) component has been determined through manual tuning of various hyperparameters.

Optimizer. Four optimizers have been tested during a 500 epochs training. Among the chosen ones, Adam is the top performer, achieving 76.17% accuracy in test phase and has the fastest convergence, closely followed by RMSProp that reaches a lower accuracy value equal to 74.72%. In contrast, Adagrad and SGD (with momentum) delivery slower convergence and lower performance of 69.08% and 68.61% respectively. In the rest of the experiments, we use the Adam optimizer.

Activation Function. Table 3 compares the results when different activation functions are considered. For each function, the training loss, test accuracy, recall, precision and F1 score values are represented. LeakyReLU and ReLU have similar behavior with the first obtaining the best overall results. We use LeakyReLU in all subsequent experiments.

Per-Class Results. Table 4 represents the F1 Score per-class. As can be seen, the Iron Cross has the highest value. This high level of accuracy is related to the fact that OpenPose can reliably detect poses for these skills, considering that Iron Cross is a vertical skill and it is not upside down, with full visible body (without limbs overlapping). One Arm Handstand, on the other hand, is the most challenging because of the upside down body position.

Accuracy at Segment Edges. We further investigated frame classification accuracy at different distances from the edges of ground truth skill segments. The histogram shown in Figure 4 illustrates that frames closer to the edges of the segments generally have a lower accuracy than those with a more centered position (further from the edges). This correlation is related to the higher human and model uncertainty when classifying frames representing a transition from a skill to another movement or vice versa.

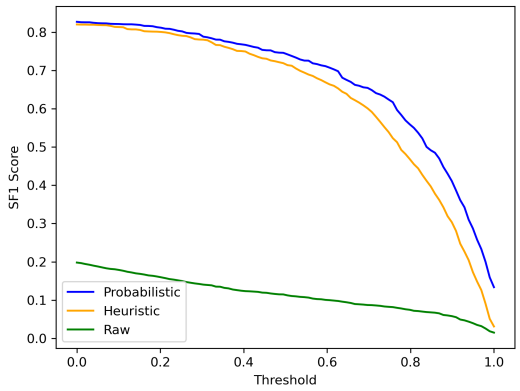

Figure 5: Threshold-SF1 curves comparing heuristic with respect to the probabilistic method.

5.2 Temporal Segmentation Algorithms

We evaluate the effectiveness of the temporal segmentation algorithms using the SF1 and ASF1 metrics considered in (Furnari et al., 2018).

The SF1 metric is a threshold-dependent, segment-based F1 measure. It is computed using precision and recall values for a specific threshold γ:

$$SF1^{(\gamma)}(t) = 2 \cdot \frac{\text{precision}^{(\gamma)}(t) \cdot \text{recall}^{(\gamma)}(t)}{\text{precision}^{(\gamma)}(t) + \text{recall}^{(\gamma)}(t)}$$

ASF1 (Average SF1) is the overall performance score of the segmentation method, computed as the average SF1 score across a set of thresholds $T = \{t \text{ such that } 0 \leq t \leq 1\}$:

$$ASF1^{(\gamma)} = \frac{\sum_{t \in T} \text{SF1}^{(\gamma)}(t)}{|T|}$$

mASF1 (mean ASF1) is the average ASF1 scores for all considered classes ($\gamma \in 0, ..., M$). It provides an overall assessment of the method's performance across different classes.

Figure 5 illustrates the SF1 scores of the algorithms at different thresholds t. Thresholds consist of 100 values comprised between 0 and 1 which influence the precision and the recall, hence the score. As observed, the heuristic algorithm performs comparably across all thresholds, displaying a steeper decline in higher thresholds compared to the probabilistic one. In both analyses, the labels predicted by the MLP without the application of any temporal segmentation algorithm achieve limited performance. Table 2 provides the mASF1 and the ASF1 scores for each individual class.

The comparisons show that both algorithms achieve similar behavior in a range of situations included in the proposed dataset. We also note that while promising results are obtained, further enhancements can be made on the proposed dataset.

Table 2: Per-class ASF1 scores and related mASF1 measures for all compared methods.

Method	mASF1	BL	FL	FLAG	IC	MAL	NONE	OAFL	OAHS	PL	VSIT
Heuristic	0.631	0.662	0.640	0.640	0.772	0.658	0.425	0.615	0.466	0.659	0.772
Probabilistic	**0.674**	**0.723**	**0.681**	**0.641**	**0.847**	**0.713**	**0.476**	**0.658**	**0.492**	**0.714**	**0.797**
Raw	0.114	0.081	0.095	0.080	0.301	0.073	0.107	0.074	0.023	0.152	0.151

Figure 6: A scheme of the completed pipeline. From left to right: a list of the performed skills with their corresponding holding times, a frame showing the OpenPose tracking, a timeline consisting of four rows: raw prediction from the MLP, heuristic and probabilistic algorithm timeline reconstruction and the ground truth timeline. On the right, the output probabilities from the MLP and the class selected by the heuristic algorithm.

Table 3: Activation functions testing results comparison.

Activation Function	TR Loss	Test Accuracy	Recall	Precision	F1 Score
LeakyReLU	0.008	**76.17%**	0.763	**0.784**	**0.767**
ReLU	0.008	76.15%	0.762	0.782	0.765
Sigmoid	0.110	76.17%	**0.765**	0.768	0.764
Tanh	**0.007**	74.49%	0.747	0.779	0.754
SiLU	0.029	74.69%	0.747	0.777	0.752

Table 4: F1 Score per class.

Skills	BL	FL	FLAG	IC	MAL	NONE	OAFL	OAHS	PL	VSIT
F1 Score	0.83	0.74	0.83	0.91	0.74	0.63	0.76	0.64	0.80	0.88

An instance of the whole pipeline applied to a video is illustrated in Figure 6. As can be observed, the set of skills present in the video is correctly identified. However, while the first, third, fourth and fifth segments are estimated with a great level of precision (small frame divergences do not affect the quality of holding time estimation), the second element (as displayed by the pose above), is not properly segmented. This issue is caused by incorrect predictions from the MLP, which cannot be rectified by the temporal segmentation algorithms. Improving the multiclass classifier could lead to better results.

6 CONCLUSIONS

In this work, we introduced a novel dataset of isometric calisthenics skills. To construct the dataset, we crawled videos and processed them using OpenPose, enabling us to retrieve the spatial coordinates of the body joints. We benchmarked a temporal segmentation approach based on per-frame classification performed with a Multilayer Perceptron and a temporal segmentation algorithm, for which we compare two versions, a heuristic-based approach and a probabilistic one. Our analysis shows promising results, though there is potential for achieving better results through different architectures. We hope that the proposed dataset will support research in video analysis for Calisthenics skills recognition.

ACKNOWLEDGEMENTS

This research has been supported by Research Program PIAno di inCEntivi per la Ricerca di Ateneo 2020/2022 (C.d.A. del 29.04.2020) — Linea di Intervento 3 “Starting Grant” - University of Catania.

REFERENCES

Ahmadalinezhad, M. and Makrehchi, M. (2020). Basketball lineup performance prediction using edge-centric multi-view network analysis. *Social Network Analysis and Mining*, 10.

Banoth, T. and Hashmi, M. F. (2022). Yolov3-sort: detection and tracking player/ball in soccer sport. *Journal of Electronic Imaging*, 32.

Cao, Z., Hidalgo, G., Simon, T., Wei, S.-E., and Sheikh, Y. (2019). Openpose: Realtime multi-person 2d pose estimation using part affinity fields.

Cheng, Y., Fan, Q., Pankanti, S., and Choudhary, A. (2014). Temporal sequence modeling for video event detection. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2235–2242.

Ding, G., Sener, F., and Yao, A. (2023). Temporal action segmentation: An analysis of modern techniques.

Fanuel, M., Yuan, X., Nam Kim, H., Qingge, L., and Roy, K. (2021). A survey on skeleton-based activity recognition using graph convolutional networks (gcn). In *12th International Symposium on Image and Signal Processing and Analysis (ISPA)*, pages 177–182.

Furnari, A., Battiato, S., and Farinella, G. M. (2018). Personal-location-based temporal segmentation of egocentric videos for lifelogging applications. *Journal of Visual Communication and Image Representation*, 52:1–12.

Garnier, P. and Gregoir, T. (2021). Evaluating soccer player: from live camera to deep reinforcement learning.

Giancola, S., Cioppa, A., Deliège, A., Magera, F., Somers, V., Kang, L., Zhou, X., Barnich, O., Vleeschouwer, C., Alahi, A., Ghanem, B., Droogenbroeck, M., Darwish, A., Maglo, A., Clapés, A., Luyts, A., Boiarov, A., Xarles, A., Orcesi, A., and Li, Z. (2022). Soccernet 2022 challenges results. pages 75–86.

Giulietti, N., Caputo, A., Chiariotti, P., and Castellini, P. (2023). Swimmernet: Underwater 2d swimmer pose estimation exploiting fully convolutional neural networks. *Sensors*, 23(4).

Hauri, S., Djuric, N., Radosavljevic, V., and Vucetic, S. (2021). Multi-modal trajectory prediction of nba players. pages 1639–1648.

Huang, W., He, S., Sun, Y., Evans, J., Song, X., Geng, T., Sun, G., and Fu, X. (2022). Open dataset recorded by single cameras for multi-player tracking in soccer scenarios. *Applied Sciences*, 12(15).

Khobdeh, S., Yamaghani, M., and Sareshkeh, S. (2023). Basketball action recognition based on the combination of yolo and a deep fuzzy lstm network. *The Journal of Supercomputing*, pages 1–26.

Kingma, D. P. and Ba, J. (2017). Adam: A method for stochastic optimization.

Koshkina, M., Pidaparthy, H., and Elder, J. H. (2021). Contrastive learning for sports video: Unsupervised player classification.

Kulkarni, K. M. and Shenoy, S. (2021). Table tennis stroke recognition using two-dimensional human pose estimation.

Liu, Y., Hafemann, L. G., Jamieson, M., and Javan, M. (2021). Detecting and matching related objects with one proposal multiple predictions.

Low, S. (2016). *Overcoming Gravity: A Systematic Approach To Gymnastics And Bodyweight Strength (Second Edition)*. Battleground Creative.

Manafifard, M., Ebadi, H., and Abrishami Moghaddam, H. (2017). A survey on player tracking in soccer videos. *Computer Vision and Image Understanding*, 159:19–46. Computer Vision in Sports.

Martin, Z., Patel, A., and Hendricks, S. (2021). Automated tackle injury risk assessment in contact-based sports – a rugby union example.

McNally, W., Walters, P., Vats, K., Wong, A., and McPhee, J. (2021). Deepdarts: Modeling keypoints as objects for automatic scorekeeping in darts using a single camera.

Munea, T. L., Jembre, Y. Z., Weldegebriel, H. T., Chen, L., Huang, C., and Yang, C. (2020). The progress of human pose estimation: A survey and taxonomy of models applied in 2d human pose estimation. *IEEE Access*, 8:133330–133348.

Murthy, P., Taetz, B., Lekhra, A., and Stricker, D. (2023). Divenet: Dive action localization and physical pose parameter extraction for high performance training. *IEEE Access*, 11:37749–37767.

Naik, B. T., Hashmi, M. F., and Bokde, N. D. (2022). A comprehensive review of computer vision in sports: Open issues, future trends and research directions. *Applied Sciences*, 12(9).

Pidaparthy, H., Dowling, M. H., and Elder, J. H. (2021). Automatic play segmentation of hockey videos. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 4580–4588.

Rahimi, A. M., Lee, K., Agarwal, A., Kwon, H., and Bhattacharyya, R. (2021). Toward improving the visual characterization of sport activities with abstracted scene graphs. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 4495–4502.

Rahmad, N., As'ari, M. A., Ibrahim, M., Sufri, N. A. J., and Rangasamy, K. (2020). *Vision Based Automated Badminton Action Recognition Using the New Local Convolutional Neural Network Extractor*.

Ramanathan, V., Huang, J., Abu-El-Haija, S., Gorban, A., Murphy, K., and Fei-Fei, L. (2016). Detecting events and key actors in multi-person videos.

Ren, B., Liu, M., Ding, R., and Liu, H. (2020). A survey on 3d skeleton-based action recognition using learning method.

Richard, A. and Gall, J. (2016). Temporal action detection using a statistical language model. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3131–3140.

Shaikh, M. B. and Chai, D. (2021). Rgb-d data-based action recognition: A review. *Sensors*, 21(12).

Spagnolo, P., Mazzeo, P. L., Leo, M., Nitti, M., Stella, E., and Distante, A. (2014). On-field testing and evalu-

ation of a goal-line technology system. *Advances in Computer Vision and Pattern Recognition*, 71:67–90.

Suryawanshi, Y., Gunjal, N., Kanorewala, B., and Patil, K. (2023). Yoga dataset: A resource for computer vision-based analysis of yoga asanas. *Data in Brief*, 48:109257.

Vinyes Mora, S. and Knottenbelt, W. J. (2017). Deep learning for domain-specific action recognition in tennis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

Wang, C. and Yan, J. (2023). A comprehensive survey of rgb-based and skeleton-based human action recognition. *IEEE Access*, 11:53880–53898.

Xiao, J., Tian, W., and Ding, L. (2023). Basketball action recognition method of deep neural network based on dynamic residual attention mechanism. *Information*, 14(1).

Yoon, Y., Hwang, H., Choi, Y., Joo, M., Oh, H., Park, I., Lee, K.-H., and Hwang, J.-H. (2019). Analyzing basketball movements and pass relationships using real-time object tracking techniques based on deep learning. *IEEE Access*, 7:56564–56576.

Yue, R., Tian, Z., and Du, S. (2022). Action recognition based on rgb and skeleton data sets: A survey. *Neurocomputing*, 512:287–306.

Zandycke, G. V., Somers, V., Istasse, M., Don, C. D., and Zambrano, D. (2022). DeepSportradar-v1: Computer vision dataset for sports understanding with high quality annotations. In *Proceedings of the 5th International ACM Workshop on Multimedia Content Analysis in Sports*. ACM.

Zhou, F., De la Torre, F., and Hodgins, J. K. (2008). Aligned cluster analysis for temporal segmentation of human motion. In *2008 8th IEEE International Conference on Automatic Face & Gesture Recognition*, pages 1–7.

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS