Temporal segmentation of actions in fencing footwork training

Filip Malawski
Institute of Computer Science
AGH University of Science and Technology
Krakow, Poland
fmal@agh.edu.pl

Marek Krupa
Institute of Computer Science
AGH University of Science and Technology
Krakow, Poland
mkrupa@agh.edu.pl

ABSTRACT

Automatic analysis of actions in sports training can provide useful feedback for athletes. Fencing is one of the sports disciplines in which the correct technique for performing actions is very important. For any practical application, temporal segmentation of movement in continuous training is crucial. In this work, we consider detecting and classifying actions in a sequence of fencing footwork exercises. We apply pose estimation to RGB videos and then we perform per-frame motion classification, using both classical machine learning and deep learning methods. Using sequences of frames with the same class we find data segments with specific actions. For evaluation, we provide extended manual labels for a fencing footwork dataset previously used in other works. Results indicate that the proposed methods are effective at detecting four footwork actions, obtaining 0.98 F1 score for recognition of action segments and 0.92 F1 score for per-frame classification. In the evaluation of our approach, we provide also a comparison with other data modalities, including depth-based pose estimation and inertial signals. Finally, we include an example of qualitative analysis of the performance of detected actions, to show how this approach can be used for training support.

Keywords

Temporal segmentation, action recognition, sports analysis, fencing, pose estimation, motion analysis.

1 INTRODUCTION

Due to recent advances, human action recognition has found applications in areas such as human-computer interaction, assisted living systems, rehabilitation support, entertainment, surveillance, and sports analysis [KF22, BNSH20]. Supporting sports training with the information provided by various devices becomes more and more popular, not only in professional but also in amateur sports. In highly technical sports disciplines, such as fencing, it is crucial to get proper feedback on exercises in order to improve the performance of different actions. While this task is typically realized by a coach, it is possible to automatically measure several motion parameters during training and provide this information to the person performing the actions. Temporal segmentation is a crucial element of motion analysis, as it enables the automatic detection of actions that can then be evaluated.

In this work, we consider temporal segmentation of actions in fencing footwork, in which a number of relevant motion parameters can be measured. We detect and classify four relevant actions in recordings of continuous training. Our goal is to obtain a solution that provides useful feedback based on RGB video data. We employ a pose-based action detection, therefore, variations in environment conditions are handled by a state-of-the-art RGB-based pose estimation algorithm, and our models need only to focus on the patterns of motion in actions. This enables us to train the models on a relatively small dataset. Since other modalities are also commonly used for similar tasks, we compare our methods on depth-based pose estimation and inertial data as well. For evaluation we obtained extended expert manual labeling for a dataset used in previous works. We also show how pose estimation and action detection can be used to obtain specific action performance parameters. In this work we: provide expert, multi-class labeling for fencing footwork dataset, compare classical and deep learning approaches for temporal segmentation using multiple modalities, propose a proof-of-concept action performance analysis.

2 RELATED WORK

Sports support is a very promising application of automatic human motion analysis. A variety of methods, with several data modalities, have been proposed in the literature to provide automated detection, classification and evaluation of actions in a variety of sports disciplines [WWB+22, PZW+22].

2.1 Data modalities

Analysis of actions in sports can be performed with multiple different modalities [SKR+23]. An obvious and popular approach is to employ RGB videos, as those are easy to acquire. Moreover, automatic analysis based on videos mimics typical workflow of a human coach. Convolutional neural networks are currently the most common approach for spatio-temporal action recognition when using RGB videos directly [CPR+21, LLZ+20]. Inertial measurement units (IMU) are small devices mounted on a person, that can measure acceleration, angular velocity and magnetic field, as well as estimate orientation based on data fusion. IMUs are widely used in action recognition, particularly in sports, as they do not suffer from occlusions, which is a relevant problem in vision based approaches. IMUs are employed, among others, for analysis of swimming [WPTM18] and combat sports [WEST19]. Another popular approach is to employ so-called skeleton data modality - an estimation of human pose, provided as coordinates of the most relevant joints. A large number of methods took advantage of skeleton data provided by the Kinect depth sensor [ZLO+16, RLDL20]. Recently, pose estimation from RGB videos has become increasingly popular and effective [BJ21, BGR+20], allowing to obtain reliable skeleton data with a typical smartphone, without the need to use a dedicated depth sensor [MJ22].

2.2 Action recognition in sports

Depending on the considered type of sports, different problems are relevant for extracting meaningful information from sports actions recordings. In team sports spatio-temporal event detection is of particular interest [YLH19], as well as tracking of players [FSY+20] and ball [YHC+19]. In analysis of individual sports the focus is more on detection, classification and evaluation of specific actions [HIK22]. Those, however, vary greatly between disciplines, therefore automatic analysis methods are often difficult to generalize. Classification of manually segmented fragments of signals including a single action was applied, among others, in tennis [SHU+22]. Automatic, temporal segmentation of actions is usually more difficult, but necessary in real-world applications. While in some sports it is sufficient to detect subsequent repetitions of the same action, e.g. in swimming [ZXZ+17], in other disciplines a variety of actions, that can occur in almost any order, must be considered for effective analysis. Fencing is one of such disciplines, as combining different techniques in rapid and unpredictable manners is an important part of tactics.

Fencing was previously analyzed in terms of footwork classification [MK18, ZWM22], bladework classification [MRPL10] and also kinematics analysis of motion [GTF08]. In this work, we consider analysis of continuous fencing footwork training. This problem was previously addressed in [Mal20], where a single action (lunge) was detected using rule-based model. In this work, using the same dataset, we extend manual labeling of data to include total of four actions (step forward, step backward, lunge, return from lunge). Next, we propose and evaluate action detection methods based on both classical machine learning and deep learning methods. Finally, we show how the proposed approach can be used to provide useful feedback to fencers.

3 FENCING FOOTWORK

Fencing training includes two main elements - footwork and bladework. Those are practiced separately in specific exercises and then jointly in combined exercises. In this work, we consider only the footwork. The main actions in footwork are steps forward and backward, as well as fencing lunge and return from the lunge. Fencers move in a sideways position (see Fig. 1 left), with the blade always pointed towards the opponent, therefore we can distinguish the front and the back leg. In fencing steps (see Fig. 1 middle), it is important to maintain proper distance between both legs, as well as correct knee bend. Fencing lunge (see Fig. 1 right) is a dynamic forward motion used during offensive actions. Proper lunge action is initiated by lifting the front foot toes, then thrusting the front leg, straightening the knee and finally landing, with knee angle in resting position at least 90 degrees. Proper return to basic fencing pose depends on not relaxing legs muscles between the lunge and the return. It is worth noting, that steps, lunge, and return have some variations, e.g. including small jump motion. In all actions, time and range of performance are also important. Tracking those parameters of performing fencing footwork exercises provides relevant feedback to a fencer, which can aid them to progress faster. Automating this process requires temporal segmentation of continuous training, as well as estimation of specific motion parameters.

4 METHODS

In this section we describe employed data and labeling process, pose estimation, temporal segmentation and performance evaluation.

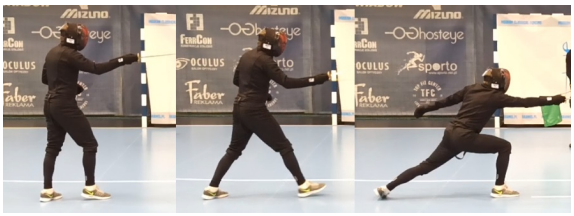

Figure 1: Fencing base pose (left), step forward (middle) and lunge (right).

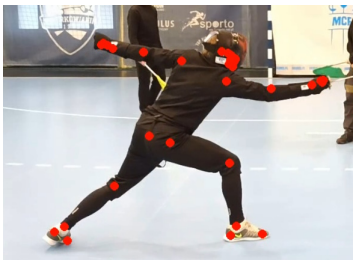

Figure 2: Fencing lunge - pose estimation (red dots indicate detected joints)

4.1 Data labeling

We employ a dataset previously used in [Mal20]. It contains recordings of continuous fencing footwork training, acquired with the Kinect sensor and two IMUs, mounted on the chest and on the elbow of the front arm. The Kinect data includes RGB video, depth video, and skeleton estimation, while the two IMUs provide acceleration, angular velocity, and magnetic field. All data are synchronized with a common sampling frequency 30Hz. Aforementioned previous work compared the detection of a single action (fencing lunge) using skeleton data estimated from depth maps and acceleration from IMUs. In this work we provide additional manual labeling of all recordings in the dataset, to include four actions: step forward, step backward, lunge, and return. Manual segmentation is performed by an expert fencer (15 years of experience). A custom tool was developed for labeling, which provided user interface for frame-by-frame viewing of the video and selecting type, start frame and end frame of each action. It is worth noting, that the expert's opinion is that the exact start and end points of actions are sometimes unclear, as the actions may partly overlap or some additional movement between actions may be present.

4.2 Pose estimation

Our goal is to obtain reliable action detection based solely on RGB videos. We use RGB pose estimation as an intermediate representation of motion, therefore variability of environment, lighting, and poses is already captured in the pose estimation model, and our models can focus on the performed actions. This allows us to obtain effective action recognition even though the dataset is relatively small (28 recordings lasting approx. 30 seconds each). We employ BlazePose model included in MediaPipe library [BGR+20]. While our main focus is on action detection from RGB videos (using pose estimation as intermediate representation), for comparison, we evaluate our methods on the depth-based skeleton and inertial modalities as well.

Detection of actions in recorded video signal starts with running the BlazePose algorithm from MediaPipe library (see Fig. 2). It provides estimation for 33 landmarks, including 11 face keypoints and 22 most relevant joints (shoulders, elbows, wrists, thumbs, pinky and index fingers, hips, knees, ankles, heels, feet index). Pose is estimated in each frame, using the fastest out of three MediaPipe models, with tracking mode enabled (detection from previous frames is used). While the other two models are larger and therefore more effective, we found that there is little practical difference in accuracy of the models, while the difference in speed is significant. The fastest model is able to run at 15 frames per second on a mid-range smartphone, which enables real-time tracking. It is worth noting, that while BlazePose is able to provide 3D estimation of joint positions, we employ only x and y coordinates, as motion along the z-axis (depth) is not relevant in this scenario.

4.3 Temporal segmentation

We perform temporal detection of actions by classifying each frame based on its context (neighboring frames). We investigate two main approaches: classical machine learning (CML) and deep learning (DL). In the CML method, we extract features in frequency domain and then we apply dimensionality reduction and classification, which is an approach proved to be effective in various motion analysis scenarios [MK18,HJ09]. We first compute per-frame x and y velocities of each joint, using the difference of their positions in neighbouring frames. Then, for each frame, we compute the discrete cosine transform (DCT) in a time window centered on this frame. The length of the window is an important hyperparameter to be selected in the experiments, as it defines the context. From the obtained DCT coefficients we remove the first one, as it corresponds to the constant component and therefore may introduce unwanted bias. DCT coefficients for x and y axes are concatened and then principal components analysis (PCA) is applied in order to remove redundant information and reduce the number of features. Finally, using obtained feature vectors, we train the support vector machine (SVM) classifier.

In the DL approach, we consider three types of neural networks: long short-term memory (LSTM), gated recurrent units (GRU), and 1-dimensional convolutional neural networks (CNN1D). Those architectures proved to be efficient for human action recognition in different applications [LWW+17, MJ22]. The first recurrent

neural network (LSTM) has two bidirectional LSTM layers with 128 units each, followed by two dense layers with 128 and 64 units. The second recurrent network (GRU), has similar architecture, with LSTM layers replaced with GRU layers. CNN1D has three 1-dimensional convolution layers with kernel size 3 and units number set to 32, 64, and 128 respectively. Between convolutions, there are max pooling operations (size 2), and at the end, global average pooling is applied, followed by a dense layer with 128 units. In all networks, there is a final layer with a size equal to the number of classes. In the case of the DL approach the input signal is also a time window of selected length, but rather than computing DCT, we use directly the sequence of joint coordinates, although filtered with Butterworth's filter, with cutoff frequency = 2Hz.

Per frame detection allows to relatively easily find action segments. One common problem that needs to be addressed when merging single frames into segments is the misclassification of single frames or even short sequences of frames during the action. In order to handle such situations, we apply postprocessing, in which short segments of frames with the same class are reclassified if they occur between two segments of another class (which becomes their new class). The maximum length of reclassified segments is set to 10, which corresponds to 330 ms. We found that such length is sufficient to remove such occurrences, while also ensuring that in this time range, there is no actual action of another class. It is also worth noting that in this work we do not address the problem of segmenting subsequent instances of the same action, e.g. multiple steps backward will be treated as a single segment of this class.

As mentioned previously, for comparison we consider also Kinect skeleton modality and IMU data. Since Kinect provides similar data as the BlazePose estimation (only with a smaller number of landmarks), the methods remain the same. IMUs provide a 3-axis measurement of acceleration, angular velocity, and magnetic field. While the nature of these data differs greatly from pose modality, these are also time signals, which can be processed in the same manner, therefore we apply the same approaches. Please note, that the methods were not optimized for the different modalities.

4.4 Performance evaluation

While the main goal of this work was temporal segmentation of actions in fencing footwork, we also include proof-of-concept methods for evaluation of performance, to provide feedback regarding the most common mistakes. We consider two motion parameters:

- Ratio of minimal feet distance to the shoulder distance during step forward and step backward actions
- Maximum angle of the front knee during the lunge action

Regarding the first motion parameter, fencing coaches recommend, that for effective moving, in forward and backward steps, fencers should keep the distance between feet similar to the distance between the shoulders. A common mistake is to have the feet too close to each other after finishing a step. Therefore, we measure minimal distance of feet in steps and compare it to the shoulder distance. Regarding the second motion parameter, the coaches state that the front leg should be fully straightened during the lunge to obtain optimal range and dynamics. Therefore, we measue the maxium knee angle in this action. Both parameters are measured using joint positions from pose estimation.

5 EXPERIMENTS

For experimental evaluation we employ dataset from [Mal20] with additionally added manual labels to include a total of 4 actions: step forward, step backward, lunge, and return, see Sec. 2.1. The dataset was acquired with 9 fencers, and for each fencer, there are 3 or 4 recordings - sequences of continuous fencing footwork training. In all experiments we employ leave-one-subject-out cross-validation, resulting in 9 folds, and the presented results are averaged from all folds. Parameters for feature extraction and classification were determined in a grid search. For the SVM approach, we used window size = 20, number of selected PCA components = 300, and regularization parameter C = 1. Neural networks (LSTM/GRU/CNN1D) were trained, respectively, on data with window size = 20/20/15 using Adam optimizer, with learning rate 0.001/0.0005/0.001 and batch size = 32/128/32, for 8/20/20 epochs. It is worth noting that the window size had the most impact on the results.

In the experiments we consider two scenarios: 1) detection of lunge action only, in order to compare with the previous method, and 2) detection of all four actions. For all experiments we measure precision, recall and F1 score. Precision is the ratio of correctly classified frames or actions of given class to all frames or actions classified as this action. Recall is the ratio of correctly classified frames or actions of given class to all actual frames or actions of this class. F1 score is a harmonic mean of precision and recall, which makes it a well balanced metric. Finally, we also present results for the evaluation of performance based on selected parameters.

5.1 Lunge detection

First, we investigate the effectiveness of detecting only the lunge action in order to compare proposed automatic approaches with the previous, rule-based method described in [Mal20]. We present both per-action and per-frame classification results. An action is considered to be detected correctly if the middle frame of the detected segment lies between the start and end frames of

the ground truth segment. We also provide results for finding the first and the last frame of action. While in some actions exact start and end points are not that important, in lunge action start point needs to be detected accurately in order to evaluate correctness in terms of relative motion of body and hand with the weapon. Finally, we depict an example of detection in a plot including ground truth and detected segments.

Modality	Method	F1	Prec.	Recall
RGB pose	SVM	1.00	1.00	1.00
	LSTM	0.99	1.00	0.99
	GRU	1.00	1.00	1.00
	CNN1D	0.97	0.99	0.96
Kinect pose	SVM	0.99	1.00	0.98
	LSTM	1.00	1.00	0.99
	GRU	0.99	1.00	0.98
	CNN1D	0.99	0.99	0.99
	Rules	1.00	1.00	1.00
IMU	SVM	0.94	0.96	0.93
	LSTM	0.91	0.95	0.88
	GRU	0.92	0.97	0.87
	CNN1D	0.83	0.94	0.74
	Rules	0.99	0.99	0.99

Table 1: Single class (lunge) per-action classification results. Results for rule-based method included from previous work [Mal20].

Modality	Method	F1	Prec.	Recall
RGB pose	SVM	0.94	0.96	0.93
	LSTM	0.90	0.94	0.88
	GRU	0.92	0.94	0.92
	CNN1D	0.90	0.94	0.88
Kinect pose	SVM	0.90	0.94	0.87
	LSTM	0.93	0.94	0.92
	GRU	0.91	0.93	0.90
	CNN1D	0.90	0.93	0.88
IMU	SVM	0.82	0.85	0.81
	LSTM	0.80	0.84	0.78
	GRU	0.80	0.86	0.77
	CNN1D	0.73	0.83	0.71

Table 2: Single class (lunge) per-frame classification results.

Our analysis of the results starts with per-action detection, as presented in Table 1. The referenced rule-based method obtained perfect detection of lunge actions using the Kinect pose estimation. The proposed method allowed us to obtain the same result using pose estimation from RGB data and either an SVM classifier or GRU neural network. Other methods also obtain very high results using both RGB and Kinect pose estimations. Interestingly, for the IMU data, learning methods are less effective than the rule-based approach. More specific features may be needed for this modality. Per-frame results (see Table 2) also indicate that RGB pose estimation combined with SVM or GRU is the most effective approach. In terms of finding exact start and end points (see Table 3), RGB pose estimation with GRU is the most accurate of the proposed methods, while still slightly less effective than the rule-based method. Finally, we can also observe proper detection of the lunge action in the plot in Fig. 3.

Modality	Method	Start err.	End err.
RGB pose	SVM	1.64 ± 2.22	0.99 ± 1.52
	LSTM	2.11 ± 1.32	0.99 ± 0.57
	GRU	1.51 ± 0.75	0.77 ± 0.44
	CNN1D	2.08 ± 2.84	0.92 ± 0.63
Kinect pose	SVM	1.86 ± 1.06	1.36 ± 0.71
	LSTM	1.42 ± 0.74	1.08 ± 0.48
	GRU	1.69 ± 0.81	1.39 ± 1.13
	CNN1D	1.69 ± 1.07	1.15 ± 0.43
	Rules	1.23 ± 1.17	0.66 ± 0.65
IMU	SVM	2.95 ± 1.69	3.51 ± 3.38
	LSTM	3.48 ± 2.63	3.39 ± 3.17
	GRU	3.18 ± 2.28	2.67 ± 1.70
	CNN1D	6.58 ± 9.17	4.97 ± 9.66
	Rules	2.57 ± 1.58	2.49 ± 1.70

Table 3: Single class (lunge) start and end frame detection error given in frames, with mean and standard deviation. Results for rule-based method included from previous work [Mal20].

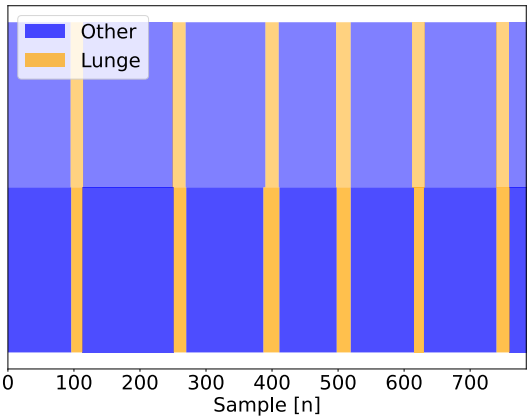

Figure 3: Example of detection of lunge action in continuous recording. Action segments are color-coded. Bottom half represents ground truth, while upper half represents detection results.

5.2 Multi-class detection

One of the key limitations of the previous rule-based method is that it does not generalize well to other actions. Defining manual rules for multiple actions is time-consuming and prone to errors. Therefore we investigate learning approaches for temporal segmentation of four actions using the extended manual labeling provided by an expert fencer.

Modality	Method	F1	Prec.	Recall
RGB pose	SVM	0.98	0.97	0.99
	LSTM	0.92	0.88	0.97
	GRU	0.96	0.94	0.98
	CNN1D	0.94	0.90	0.98
Kinect pose	SVM	0.92	0.88	0.97
	LSTM	0.97	0.95	0.99
	GRU	0.96	0.94	0.98
	CNN1D	0.94	0.91	0.97
IMU	SVM	0.87	0.82	0.93
	LSTM	0.88	0.82	0.94
	GRU	0.87	0.81	0.93
	CNN1D	0.79	0.72	0.89

Table 4: Multi class per-action classification results.

Modality	Method	F1	Prec.	Recall
RGB pose	SVM	0.92	0.92	0.92
	LSTM	0.87	0.87	0.87
	GRU	0.90	0.90	0.90
	CNN1D	0.88	0.88	0.88
Kinect pose	SVM	0.87	0.87	0.87
	LSTM	0.89	0.89	0.89
	GRU	0.89	0.89	0.89
	CNN1D	0.88	0.88	0.88
IMU	SVM	0.75	0.75	0.75
	LSTM	0.75	0.75	0.75
	GRU	0.72	0.72	0.72
	CNN1D	0.66	0.66	0.66

Table 5: Multi class per-frame classification results.

Results in Table 4 indicate that the most effective approach for detection of multiple actions is the SVM classifier applied to pose estimation from RGB video, as it obtained F1 score = 0.98, precision = 0.97 and recall = 0.99. GRU network is a close second for this modality with F1 score = 0.96, precision = 0.94, and recall = 0.98. IMU data provides significantly less accurate detection, with the best F1 score = 0.88 obtained with LSTM neural network. As mentioned before, the

Modality	Method	Start err.	End err.
RGB pose	SVM	1.47 ± 1.15	1.48 ± 1.24
	LSTM	2.11 ± 0.96	2.33 ± 1.28
	GRU	1.74 ± 0.98	1.72 ± 0.79
	CNN1D	1.99 ± 0.95	1.94 ± 1.06
Kinect pose	SVM	2.36 ± 0.78	2.37 ± 0.63
	LSTM	1.77 ± 0.66	1.73 ± 0.67
	GRU	1.87 ± 0.85	2.05 ± 1.18
	CNN1D	1.93 ± 0.78	1.83 ± 0.67
IMU	SVM	3.96 ± 2.83	4.32 ± 3.61
	LSTM	4.34 ± 1.25	4.74 ± 1.72
	GRU	4.12 ± 1.07	4.76 ± 1.29
	CNN1D	4.44 ± 2.61	4.51 ± 2.72

Table 6: Multi class start and end frame detection error given in frames (including mean and standard deviation).

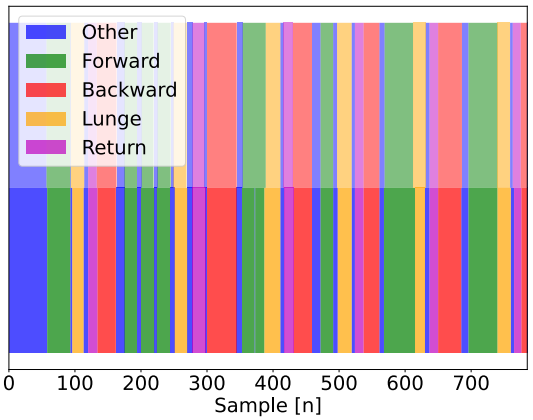

Figure 4: Example of detection of four actions in continuous recording. Action segments are color-coded. Bottom half represents ground truth, while upper half represents detection results.

proposed methods were not optimized for this modality, which may be the reason for lower effectiveness. Results for per-frame detection (see Table 5) indicate the same methods to be the most effective, but also show that for this application RGB pose estimation provides more relevant information than Kinect skeleton data. Error in detecting the start frame (see Table 6) is lower than in the case of a single action, however, end frame error is higher. Both start and end frame errors correspond to approx. 50 ms, which is sufficient for performance analysis. For a visual representation of multi-class temporal segmentation see the plot depicted in Fig. 4.

5.3 Action performance evaluation

While the main goal of this work was to perform temporal segmentation of actions in fencing footwork, we also include a limited action performance analysis in order to show the potential of the final application of the proposed methods. In the detected step forward actions we measure the ratio of the minimum distance of feet to the distance of shoulders. Fig. 5 presents an example of correct (left) and incorrect (right) poses in terms of feet distance. The ratio parameter for the depicted correct pose is 0.96, while for the incorrect pose, it is 0.84. Recommendation from a fencing coach is that the ratio should be close to 1. In Fig. 6 correct lunge is the one with a straight front leg (left image), while the incorrect is the one with a bent knee (right image). The knee angle computed using pose estimation is 177 degrees and 156 degrees respectively. Expected angle for a correct action is approx. 180 degrees (straight leg). As we can see, by using detected actions and dependencies between joints in pose estimation, we can find occurrences of incorrectly performed actions and therefore provide useful feedback to the fencer.

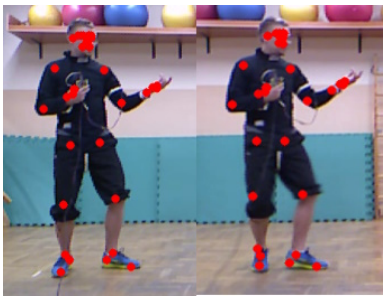

Figure 5: Example of performance analysis in step action. Correct distance between feet (right) vs incorrect (left).

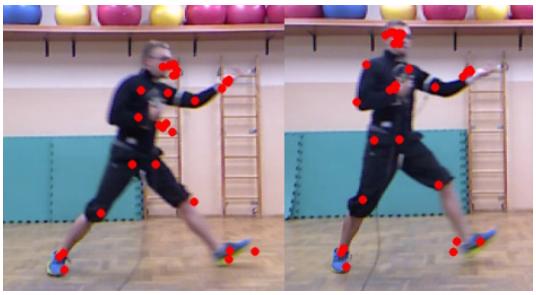

Figure 6: Example of performance analysis in lunge action. Correct knee angle (right) vs incorrect (left).

6 CONCLUSIONS

In this work, we proposed methods for support of fencing footwork exercises. Our approach employs pose estimation from RGB videos, which does not require any devices other than a smartphone, as opposed to previously proposed methods which relied on depth cameras and inertial sensors. This facilitates introducing proposed solution to fencing trainings. We evaluated classical and deep learning methods for the task. Both approaches yielded similar results, with SVM performing slightly better than best neural network architecture (GRU). We expect, that deep learning approaches would be more effective with a larger dataset. Overall, obtained results are very good. Considering best obtained multi-class action classification F1 score = 0.98, proposed method could be used in practical application. Detection of start and end frames, relevant for some actions, is also accurate - average error 1.47 frame and 1.48 frame respectively, which corresponds roughly to 50ms. Also, the proof-of-concept action performance analysis produced promising results, even though it requires more thorough evaluation, for which additional, specific data is needed.

Future works can be realized in multiple directions. First of all, additional, less common footwork actions can be added, such as dodging. Secondly, additional segmentation of sequences of the same actions (e.g. multiple steps forward or backward) can be considered. Moreover, automatic analysis of bladework would be beneficial for the fencers as well, even though it may prove more difficult to realize. Fusion of visual and inertial data may be useful in this regard. Finally, more qualitative motion parameters can be extracted for the analyzed actions, therefore providing the fencers with additional relevant feedback. However, evaluation of qualitative analysis will require recording additional data with actions performed correctly and incorrectly. Is it also worth noting, that the proposed methods could be used for real-time analysis, which may be used to deliver feedback while training, rather than only when viewing a recording. Such feedback could be delivered e.g. by generating sounds, visual signals or even spoken comments.

7 ACKNOWLEDGMENTS

The research presented in this paper was supported by the National Centre for Research and Development (NCBiR) under Grant No. LIDER/37/0198/L-12/20/NCBR/2021. We also thank Aramis Fencing School (aramis.pl) for providing experts' consultations.

8 REFERENCES

[BGR+20] Valentin Bazarevsky, Ivan Grishchenko, Karthik Raveendran, Tyler Zhu, Fan Zhang, and Matthias Grundmann. Blazepose: On-device real-time body pose tracking. *arXiv preprint arXiv:2006.10204*, 2020.

[BJ21] Rishabh Bajpai and Deepak Joshi. Movenet: A deep neural network for joint profile prediction across variable walking speeds and slopes. *IEEE Transactions on Instrumentation and Measurement*, 70:1–11, 2021.

[BNSH20] Djamila Romaissa Beddiar, Brahim Nini, Mohammad Sabokrou, and Abdenour Hadid. Vision-based human activity recognition: a survey. *Multimedia Tools and Applications*, 79:30509–30555, 2020.

[CPR+21] Chun-Fu Richard Chen, Rameswar Panda, Kandan Ramakrishnan, Rogerio Feris, John Cohn, Aude Oliva, and Quanfu Fan. Deep analysis of cnn-based spatio-temporal representations for action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6165–6175, 2021.

[FSY+20] Na Feng, Zikai Song, Junqing Yu, Yi-Ping Phoebe Chen, Yizhu Zhao, Yunfeng He, and Tao Guan. Sset: a dataset for shot segmentation, event detection, player tracking in soccer videos. *Multimedia Tools and Applications*, 79:28971–28992, 2020.

[GTF08] M Gholipour, A Tabrizi, and F Farahmand. Kinematics analysis of lunge fencing using stereophotogrametry. *World Journal of Sport Sciences*, 1(1):32–37, 2008.

[HIK22] Kristina Host and Marina Ivašić-Kos. An overview of human action recognition in

sports based on computer vision. *Heliyon*, 8(6):e09633, 2022.

[HJ09] Zhenyu He and Lianwen Jin. Activity recognition from acceleration data based on discrete consine transform and svm. In *2009 IEEE international conference on systems, man and cybernetics*, pages 5041–5044. IEEE, 2009.

[KF22] Yu Kong and Yun Fu. Human action recognition and prediction: A survey. *International Journal of Computer Vision*, 130(5):1366–1401, 2022.

[LLZ+20] Jun Li, Xianglong Liu, Wenxuan Zhang, Mingyuan Zhang, Jingkuan Song, and Nicu Sebe. Spatio-temporal attention networks for action recognition and detection. *IEEE Transactions on Multimedia*, 22(11):2990–3001, 2020.

[LWW+17] Chuankun Li, Pichao Wang, Shuang Wang, Yonghong Hou, and Wanqing Li. Skeleton-based action recognition using lstm and cnn. In *2017 IEEE International conference on multimedia & expo workshops (ICMEW)*, pages 585–590. IEEE, 2017.

[Mal20] Filip Malawski. Depth versus inertial sensors in real-time sports analysis: A case study on fencing. *IEEE sensors journal*, 21(4):5133–5142, 2020.

[MJ22] Filip Malawski and Bartosz Jankowski. Depth-based vs. color-based pose estimation in human action recognition. In *Advances in Visual Computing: 17th International Symposium, ISVC 2022, San Diego, CA, USA, October 3–5, 2022, Proceedings, Part I*, pages 336–346. Springer, 2022.

[MK18] Filip Malawski and Bogdan Kwolek. Recognition of action dynamics in fencing using multimodal cues. *Image and Vision Computing*, 75:1–10, 2018.

[MRPL10] G Mantovani, A Ravaschio, P Piaggi, and A Landi. Fine classification of complex motion pattern in fencing. *Procedia Engineering*, 2(2):3423–3428, 2010.

[PZW+22] Yiqun Pang, Changnian Zhang, Yibing Wang, Qiurui Wang, and Mingyang Wang. Analysis of computer vision applied in martial arts. In *2022 2nd International Conference on Consumer Electronics and Computer Engineering (ICCECE)*, pages 191–196. IEEE, 2022.

[RLDL20] Bin Ren, Mengyuan Liu, Runwei Ding, and Hong Liu. A survey on 3d skeleton-based action recognition using learning method. *arXiv preprint arXiv:2002.05907*, 2020.

[SHU+22] Anik Sen, Syed Md Minhaz Hossain, Russo-MohammadAshraf Uddin, Kaushik Deb, and Kang-Hyun Jo. Sequence recognition of indoor tennis actions using transfer learning and long short-term memory. In *Frontiers of Computer Vision: 28th International Workshop, IW-FCV 2022, Hiroshima, Japan, February 21–22, 2022, Revised Selected Papers*, pages 312–324. Springer, 2022.

[SKR+23] Zehua Sun, Qiuhong Ke, Hossein Rahmani, Mohammed Bennamoun, Gang Wang, and Jun Liu. Human action recognition from various data modalities: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3200–3225, 2023.

[WEST19] Matthew TO Worsey, Hugo G Espinosa, Jonathan B Shepherd, and David V Thiel. Inertial sensors for performance analysis in combat sports: A systematic review. *Sports*, 7(1):28, 2019.

[WPTM18] Matthew TO Worsey, Rebecca Pahl, David V Thiel, and Peter D Milburn. A comparison of computational methods to determine intrastroke velocity in swimming using imus. *IEEE Sensors Letters*, 2(1):1–4, 2018.

[WWB+22] Fei Wu, Qingzhong Wang, Jiang Bian, Ning Ding, Feixiang Lu, Jun Cheng, Dejing Dou, and Haoyi Xiong. A survey on video action recognition in sports: Datasets, methods and applications. *IEEE Transactions on Multimedia*, pages 1–25, 2022.

[YHC+19] Young Yoon, Heesu Hwang, Yongjun Choi, Minbeom Joo, Hyeyoon Oh, Insun Park, Keon-Hee Lee, and Jin-Ha Hwang. Analyzing basketball movements and pass relationships using realtime object tracking techniques based on deep learning. *IEEE Access*, 7:56564–56576, 2019.

[YLH19] Junqing Yu, Aiping Lei, and Yangliu Hu. Soccer video event detection based on deep learning. In *MultiMedia Modeling: 25th International Conference, MMM 2019, Thessaloniki, Greece, January 8–11, 2019, Proceedings, Part II 25*, pages 377–389. Springer, 2019.

[ZLO+16] Jing Zhang, Wanqing Li, Philip O Ogunbona, Pichao Wang, and Chang Tang. Rgb-d-based action recognition datasets: A survey. *Pattern Recognition*, 60:86–105, 2016.

[ZWM22] Kevin Zhu, Alexander Wong, and John McPhee. Fencenet: Fine-grained footwork recognition in fencing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3589–3598, 2022.

[ZXZ+17] Zhendong Zhang, Dongfang Xu, Zhihao Zhou, Jingeng Mai, Zhongkai He, and Qining Wang. Imu-based underwater sensing system for swimming stroke classification and motion analysis. In *2017 IEEE International Conference on Cyborg and Bionic Systems (CBS)*, pages 268–272. IEEE, 2017.