Recognition of Action Dynamics in Fencing Using Multimodal Cues

Filip Malawski, Bogdan Kwolek*

AGH University of Science and Technology, 30 Mickiewicza Av., 30-059 Krakow, Poland

Abstract

Most current approaches to action recognition follow strategies, which permit classification of significantly different actions. However, in some sports disciplines, actions may be distinguished mainly by the dynamics of the motion rather than the trajectory. In this work, we propose a novel approach for recognition of sports actions. The novelty consists in the use of dynamics in the analysis of similar motion patterns. We propose informative motion descriptors based on accelerometric data, skeleton joints features and depth maps, and demonstrate their potential to model the motion dynamics. We show that fusing data from multiple modalities permits better recognition accuracy. We make publicly available a dedicated dataset with fencing footwork samples of ten fencers that consists of depth, skeletal and inertial data of six types of dynamic actions, most of which have similar average trajectories but different dynamics of the motion. We show that on our Fencing Footwork Dataset the proposed method outperforms current state-of-the-art methods for general action recognition.

Keywords: Action recognition, depth maps, multimodal cues, motion analysis

1. Introduction

Due to highly competitive nature of sports, athletes and coaches are eager to adapt and practically verify new training methods as well as innovative technologies for training support. The use of technology in sports has enhanced sports facilities and equipment design in a wide range of sports disciplines. Biomechanical analysis of sports movements allows better understanding how the human body behaves during various sports actions and subsequently develop better training methods as well as reduce injury risk [1]. Each sports discipline developed a set of exercises aimed at perfecting particular skills. Employing modern technologies permits the athletes to master these skills in shorter time and increase their overall performance [2].

Currently, an analysis of sports actions begins to play a crucial role in the training process in several disciplines [3]. Its importance comes from the possibility of providing relevant feedback. Recognition of actions is one of the basic issues that such analysis tools need to address. This includes both general identification of different sports activities [4] as well as discipline–specific actions, for instance, swim strokes [5]. Action recognition (AR) is an important research topic with many possible applications in various areas [6]. Thus, multiple methods have already been proposed [7]. However, literature devoted to both general and sports action recognition consider only detection and classification of significantly different actions from each other [8]. Even though some of the popular AR databases contain actions, which are similar to each other to some extent [9], they all have distinct trajectories and are easily recognizable for humans. In some sports disciplines, however, actions may be distinguished mainly by the dynamics of the motion rather than the trajectory. These actions may be difficult to distinguish even for a human, especially for a person not acquainted with a particular sports discipline.

In this paper, we propose a novel approach for recognition of sports actions. The novelty consists in the use of dynamics in the analysis of the similar motion patterns. We propose informative motion descriptors and demonstrate that they model the dynamics better, and that the classifiers using them achieve superior results in comparison to state–of–the–art algorithms. We propose a method for recognition of action dynamics, which is based on accelerometric data, skeleton joints features and a novel depth–based descriptor, as well as employs

*Corresponding author
URL: http.agh.edu.pl/~bkw (Bogdan Kwolek)

a neural network for fusion of the multimodal data features. We consider 6 types of dynamic actions, which include 4 types of fencing lunge each having a very similar average trajectory, but considerably different dynamics of the motion. We make publicly available a dedicated dataset with fencing footwork samples of 10 fencers, including depth, skeletal and inertial data. We show that on our fencing footwork dataset the proposed method outperforms current state–of–the–art methods for general action recognition.

2. Related Work

Analysis of actions in sports has been addressed by several researchers, who investigated several types of sensors. Color cameras are used for visual tracking of players in beach volleyball [10], interactive feedback in martial arts [11] as well as for analysis of golf swing [12]. Depth sensors, such as the Kinect sensor, have become quite popular recently, as they are robust to illumination changes and provide 3D information. Applications of depth cameras include recognition of karate techniques [13], tracking of golf swings [14] and analysis of tennis shots [15]. The Kinect provides not only depth data but also skeleton data, which serves as a low–cost motion capture system [16, 17]. However, some applications require higher measurement precision, therefore more advanced motion capture systems are also used [18]. Inertial measurement units (IMUs) are commonly used as well [19], since they provide high sampling frequency and permit analysis of movements, whose distinguishing may be difficult on the basis of color or depth data. IMUs are employed in disciplines such as golf [20], dressage riding [21] or swimming [5]. Fusion of data from multiple modalities is employed quite rarely, although some recent works in this area can be found [22, 23].

Most of the aforementioned work employs custom datasets. There are a few publicly available datasets for sports action recognition: THETIS Three Dimensional Tennis Shots Dataset, which consists of 12 types of tennis actions recorded with the Kinect [15]; UCF Sports Dataset [8], which includes color videos of 10 different sports activities, such as diving, kicking or riding horse; G3D Gaming Action Dataset [24], with Kinect recordings of 20 gaming actions such as walk, run, jump, tennis serve, golf swing, throw bowling ball, etc. For general action recognition a larger variety of datasets is available. The most frequently used benchmark datasets are: KTH dataset (color videos, 6 classes) [25]; Weizmann dataset (color videos, 10 classes) [26]; MSR Action 3D dataset (Kinect depth data, 20 classes) [9]; MSR Daily Activity 3D dataset (Kinect depth data, 16 classes) [27]. Some of these datasets include sports related actions such as running, kicking, throwing, tennis serve, boxing, etc. Nevertheless, the available datasets, both for sports and general action recognition, lack examples of actions, which differ in the dynamics of the movement rather than in the trajectory. Motivated by the importance of dynamics in analysis of motions in sports in one hand, and from the other hand by the lack of datasets targeting this very important issue, we propose a challenging Fencing Footwork Dataset (FFD) and make it publicly available to stimulate further research in this area.

The existing work in the area of fencing includes analysis of lunge performance and biomechanics with: surface electromyography and high speed cameras [28], frontal and lateral videos [29] and stereophotogrammetry [30]. Classification of weapon actions with motion capture system is presented in [31]. However, as mentioned previously, there are no publicly available datasets for fencing analysis. Moreover, no significant work on analysis of dynamics of different types of lunges exists.

So far, the handcrafted features for action description and representation have achieved significant performance on a variety of action recognition benchmarks and datasets [7, 32]. In [33], taking into account object relationships in action representation, two types of motion reference points are examined to alleviate the effect of camera movement, which frequently takes place in unconstrained environments. A 3D convolutional neural network (C3D) [34] that has been trained on a large–scale video dataset is capable of modeling appearance and motion information simultaneously. The network has been trained on Sports–1M dataset [35], which consists of 1 million YouTube videos belonging to a taxonomy of 487 classes of sports. In this work, we decided not to use RGB images for several reasons:

- Lighting conditions – Fencers usually train in large rooms, where lighting is often not sufficient for even a high-end consumer RGB camera to provide images without blur, particularly in the case of fast, dynamic movement. For classification of similar actions, dealing with blurred data is crucial. Depth data provides resistance in such situations.

- Background – Since the training usually takes place in large rooms, multiple moving persons are expected to be present in the background, usually all dressed in single–color uniforms (white or black), which makes it extremely hard for most

RGB–based algorithms to extract motion information.

- Privacy – Contrary to the depth data, RGB videos allow for easy person identification. It is well known that systems that preserve privacy are much more acceptable.

We show experimental results that were achieved on our dataset by the C3D algorithm. Our algorithm does not use RGB data since our intention was to develop an illumination invariant algorithm, which will be capable of working in poor illumination conditions. Thus, we focused on person silhouettes extracted at low computational cost by the Kinect sensor as well as precise joint positions.

In addition to RGB videos, depth maps have been successfully employed in many approaches for action recognition [36]. The methods proposed in [37, 38] achieve competitive results on many datasets and are frequently used in evaluation of depth–based algorithms. As shown in recent surveys [6, 39], the learned representations have considerable potential in action recognition. In [40], on the basis of skeleton sequence, a set of vectors is generated with the selected pairs of joints for every body part. Afterwards, feature arrays are transformed into gray images, which are finally fed to a deep learning architecture. The method achieves superior results on three datasets and outperforms previous methods.

3. Fencing Footwork Dataset

At the beginning of this section we discuss basic footwork in fencing. Afterwards, we present Fencing Footwork Dataset.

3.1. Footwork in Fencing

Fencers move in a sideways position, in a straight line, approaching or moving away from the opponent. In a basic position the knees are slightly bend and the armed hand is directed towards the competitor, see Figure 1 left. Due to the sideways position we can distinguish the front and back leg. Basic footwork actions include steps and lunges. A step forward is initiated by the front leg, and then followed by the back leg, therefore each step is finished in the basic position. Step backward is similar, but started with the back leg. Lunge allows to dynamically shorten the distance to the opponent during an offensive action. It is performed by first slightly lifting the front leg and then pushing off with the back leg. Resulting position is pictured in Figure 1 right. Lunge is usually finished by bending the knee of the back leg, pushing off backwards with the front leg and therefore returning to the basic position.

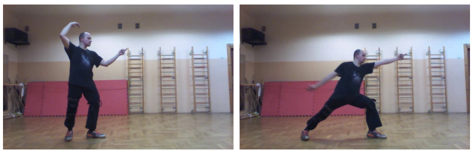

Figure 1: Fencing position (left) and fencing lunge (right).

According to prof. Czajkowski, one of the inventors of modern theory of fencing, there are four basic types of lunges, which vary in the dynamics of the motion [41]:

- rapid – very fast, performed in relatively short distances, intended for surprise attacks
- with increasing speed – slow at the beginning, but accelerated during the action, useful for feint attacks
- with waiting – with a short pause in the first stage of the lunge, during which the fencer observes the reaction of the opponent and performs a counter–action if necessary
- jumping – sliding – longest type of lunge, the fencer jumps forward with the front leg, while the back leg slides on the floor, intended for complex offensive actions

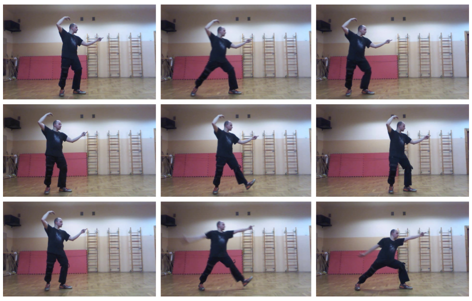

Figure 2: Key poses (beginning, middle, end) of fencing actions: top row – step backward, middle row – step forward, bottom row – lunge.

While distinguishing between step forward, step backward and lunge is rather straightforward, distinguishing between different types of lunge is not easy, as they vary mostly in the dynamics of the motion. Moreover, performance of the different lunge actions varies slightly between fencers, as it is influenced by their physical capabilities, such as speed or flexibility, as well as their skills. For this reason, recognition of different types of lunge may be a difficult task, even for human, and automatic recognition of such actions is challenging. Figure 2 presents key poses of the step forward, step backward and lunge actions. Different types of lunges are not depicted, since the difference in the dynamics of the movement is difficult to illustrate with only a few static frames.

3.2. Dataset

The main focus of our research was recognizing the dynamics of movement in the context of lunge actions, nevertheless in order to verify applicability of the proposed method for identification of more distinctive actions, we consider the steps actions as well. Other fencing footwork actions are possible, although we did not find them to be relevant in this case. The Fencing Footwork Dataset includes six dynamic actions:

- rapid lunge (R),
- incremental speed lunge (IS),
- lunge with waiting (W/W),
- jumping-sliding lunge (JS),
- step forward (SF),
- step backward (SB).

The data were recorded thanks to the courtesy of Aramis Fencing School[1], one of the biggest fencing institutions in Poland. Ten fencers in total, ranging from intermediate to professional level, both male and female participated in the recording session. We recorded color, depth and skeleton data using the Kinect as well as data from x–IMU sensor [42]. The fencers were asked to attach the inertial sensor to the knee of the front leg and perform actions in a given distance (approx. 3 m) from the Kinect sensor. The actions were performed on a command, each action was repeated 10 – 11 times and each repetition was recorded as a separate data sample consisting of multiple files.

The Kinect acquired 640 × 480 16 bit depth data at 30 Hz together with automatically extracted person and skeleton data for 20 tracked joints. Figure 2 presents the different modalities for the lunge action recorded by the Kinect (color data is for illustrative purposes only and it is not included in the dataset). The x–IMU sensor operated at 256 Hz and provided 9 axis inertial data from accelerometer, gyroscope and magnetometer as well as orientation data. Depth data was recorded as an uncompressed video file, whereas skeleton and inertial data were stored in Matlab format files. The dataset is publicly available, see `http://home.agh.edu.pl/~fmal/ffd/` for more information and example data.

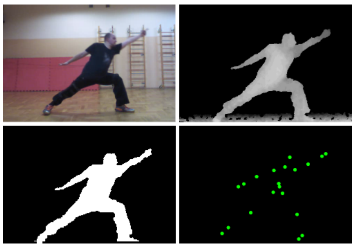

Figure 3: Lunge action recorded by the Kinect: color data (top-left), depth data (top-right), extracted silhouette (bottom left), skeleton data (bottom right).

4. Features

In order to properly describe complex fencing actions, and particularly to distinguish between similar motion patterns that accompany different fencing lunges, we developed a number of features describing the dynamics of the fencing actions. The features were verified separately in recognition of the actions as well as fused together in order to improve the recognition accuracy. At the beginning of this section we propose acceleration based features, then we discuss Joint Dynamics features that employ skeleton data, and finally we introduce a novel skeleton based–descriptor, called Local Trace Images.

4.1. Accelerometric Features

Over the last decade, a considerable number of different approaches for extracting features from accelerometer data have been proposed in the literature [43]. Those include time domain features (e.g. mean, standard deviation, variance), frequency domain features (Fourier Transform, Discrete Cosine Transform) and others such as wavelets (e.g. Haar wavelets, or Daubechies wavelets). Other data provided by IMUs, namely magnetic and gyroscope are sometimes employed as well [44]. In our research, we considered several combinations of inertial data and we found that time domain features extracted from the acceleration data are best suited for dynamics analysis [45]. In particular, magnetic and gyroscope data, which are usually used for reconstruction of trajectories [46], are less relevant for analysis of dynamics in sport actions.

[1] aramis.pl

The data acquired by accelerometer were preprocessed and then utilized by a feature extraction algorithm. First, we performed interpolation of the signal in order to ensure equal lengths of the samples. Since each sample was about 2 seconds long and the sampling frequency was set to 256 Hz, we interpolated the samples to a common length of 512 data points. In the next step we divided the samples into equal–size segments with 50% overlap. We experimented with different segment sizes (32, 64, 128, 256) and choose the final segment size of 128, which results in 7 segments per data sample, see Fig. 4.

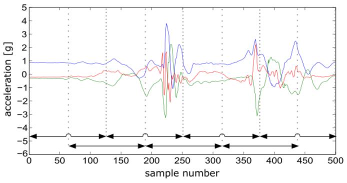

Figure 4: Sample accelerometric signal (3 axes) divided into 7 overlapping time segments.

A highpass filter was applied, with stopband frequency equal to 0.4 and passband frequency equal to 0.8 normalized frequency units. Finally, using the filtered signal, the difference between the original and the filtered signal and the first derivative of the filtered signal, we computed the following features in each segment: mean value for each axis, root mean square (RMS) value for each axis, mean value of magnitude, RMS of magnitude. This resulted in total of 24 features per segment. Considering the splitting of each sample to 7 segments, the feature vector length was equal to 168.

4.2. *Joint Dynamics Features*

In algorithms devoted to general action recognition, positions or relative positions of selected joints, both in space and time are often utilized [27, 38]. Contrary to that, in this work we employ first and second derivative of the joint positions in order to take into account velocity and acceleration patterns. The resulting motion and dynamics descriptors are called Joint Dynamics (JD) features. Similarly to processing accelerometric data the features are also extracted in data segments, although in this case, we consider multiple joints and multiple windows, as well as frequency domain rather than time domain features.

For the computation of the features we use only the eight lower body joints, namely: hips, knees, ankles and feet, which are the most relevant for the fencing footwork analysis. Having on regard that the Kinect acquires depth maps with 30 Hz and data samples are around two seconds long, we interpolate data from each joint to 64 data points. Afterwards, in order to perform multi–level analysis we divide each data sample into overlapping windows. At each level we use windows with a given size and 50% overlap. Experiments demonstrated that the best results are achieved using 3 levels with window sizes of 64, 32 and 16, which correspond to 1, 3 and 7 windows in each level, see also see Fig. 5. Such an approach allows us to capture both global and local (in terms of time) motion patterns. In each window we apply Short Time Fourier Transform (STFT) and construct a feature vector by taking absolute values of the first 3 coefficients. We consider velocity and acceleration in vertical and horizontal directions (changes in depth are irrelevant in this case), which results in four values per data point. Since we use eleven windows for each of the eight joints, the feature vector size is equal to 1056. For the configuration with six joints (hips, knees and feet), the size of the feature vector is equal to 792.

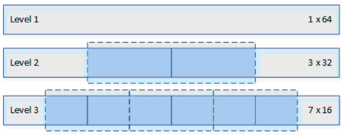

Figure 5: Multi-level windows employed for signal analysis in Joint Dynamics features.

4.3. *Local Trace Images*

Depth data has been employed in many approaches for action recognition [36]. Our algorithm, which we call Local Trace Images (LTI) is based on creating probabilistic images of motion patterns for each joint separately by employing both depth and skeleton data. Images of motion patterns were introduced with the Motion History Images (MHI) method [47], where pixel intensity represented temporal history of that point with regard to a given decay operator. This resulted in more recently moved pixels being brighter. Lately, the MHIs were used for detection of cyclic actions in indoor exercises [48]. In [27] authors use the skeleton data as an additional context for the depth data in order to compute 3D histograms around joints, which is called Local Occupancy Patterns (LOP).

For dynamics analysis we need to consider changes in motion during the whole movement. Therefore, in order to adequately model such changes in motion we propose a Trace Image descriptor. By superimposing binary silhouettes of the extracted persons without the

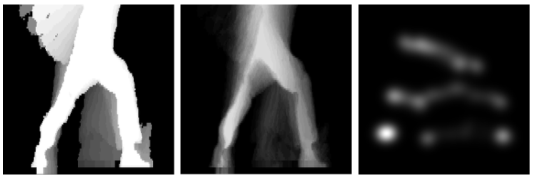

Figure 6: Lunge actions features: MHI (left), energy image (middle), Trace Image (right).

decay function used in MHI we can create energy images, which capture dynamics of the movement, see also middle image on Fig. 6, where brighter pixels represent slower motion. However, using whole silhouettes introduces significant amount of noise. Therefore, instead of using the silhouettes we model the positions of the selected joints by a two–dimensional normal distribution:

$$b = f(x, \mu, \sigma) = \frac{1}{\sqrt{2\sigma^2\pi}} \exp - \frac{(x-\mu)^2}{2\sigma^2} \qquad (1)$$

where b denotes pixel brightness in a particular frame as a function of distance from the position μ of the joint, whereas σ denotes the variance. A superposition of such Gaussians from multiple frames results in a single image, which express spatio–temporal patterns of the joints movement. We call such an image the Trace Image, see also right image on Fig. 6. However, some relevant information can be lost in some cases, mainly due to overlaps of the patterns of different joints. Therefore, we generate a separate image for each joint, which is called Local Trace Image. For each considered joint the Gaussians are first generated on a larger image, and afterwards the superposition of the relevant area (minimal square area containing only non–zero pixels) is selected as a single Local Trace Image for the particular joint. These images are then resized to a common size. In evaluations of the algorithm we found that 16×16 pixel size gives the best results. The images are then concatenated and serve as a single feature vector, see also Fig. 7.

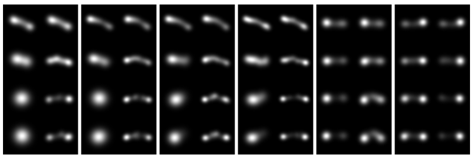

Figure 7: Local Trace Images for six fencing footwork actions. From left to right: rapid lunge (R), incremental speed lunge (IS), lunge with waiting (WW), jumping-sliding lunge (JS), step forward (SF), step backward (SB). Image for each action is a concatenation of 8 LTI, each for separate joint (from top to bottom: hips, knees, ankles, feet).

5. Classification and Fusion

Initially we classified the acceleration data using the Dynamic Time Warping (DTW) [49], which is a typical approach for time–series matching. It finds an optimal alignment between two time–series by employing dynamic programming. It allowed us to obtain promising classification accuracies for a single person [45]. However, we found it to be less suitable for person independent recognition of basic footwork in fencing. The Support Vector Machine (SVM) operating on the discussed above acceleration features achieves better results. The experiments demonstrate that it outperforms Random Forests (RF), which were evaluated in terms of classification accuracy using different number of trees grown and the number of predictors randomly tried at each split. We also employed the SVMs for the Joint Dynamics and Local Trace Images features. In evaluations of the SVM classification performance we employed both linear kernel as well as Radial Basis Function (RBF) kernel, which is commonly utilized to handle non–linear decision surfaces. Both margin parameter c as well as kernel width parameter γ were fine–tuned during the training stage.

We trained a separate SVM model for each of the feature sets (acceleration features, JD, LTI) and examined the recognition accuracy of each model. We observed that, even when the classification accuracy was similar, confusion matrices indicated that different types of features are best suited for recognizing different types of lunge. Therefore, in order to improve the classification accuracy we perform fusion of the features. The multi–class SVM can also generate probabilities for each of the output classes rather than just provide a single class label [50]. By using all 3 feature sets and their SVM models we obtained 18 probabilities, which served as input for the fusion stage. An important aspect of this approach is that regardless of the number of features in each feature set, they all contribute equal number of features to the next stage. Having on regard that the number of features is small, a Multilayer Perceptron (MLP) has been employed to fuse the multimodal features. We employed a single hidden layer, while number of neurons in the hidden layer, learning rate and momentum were used as tuning parameters. Figure 8. illustrates the entire classification process (including extraction of the features).

6. Experimental results

We validated the proposed algorithm on the Fencing Footwork Dataset. In the experiments we considered

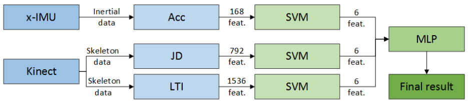

Figure 8: Block-diagram of the system. Data from the x-IMU and the Kinect sensors is used to extract 3 sets of features: acceleration (Acc), Joint Dynamics (JD), and Local Trace Images (LTI). Each feature set has a separate SVM model. Fusion of the results from SVMs is performed by the Multilayer Perceptron (MLP), which provides the final result.

Table 1: Accuracy [%] of recognition using the acceleration features (Acc) on the Fencing Footwork Dataset.

method	PD	PI
DTW	98.18	56.75
SVM linear	93.88	70.71
SVM-RBF	94.21	70.71

two cases: person dependent (PD) and person independent (PI). In the PD case the classification performance was determined using five–fold cross–validation separately for each person, where 80% of data recorded with the particular fencer has been employed for training and the other 20% of data for the test. In the PI case we performed ten–fold leave–one–out cross–validation, where in each fold nine persons were used for training and the remaining one for tests. The PD allows us to determine whether the actions of each particular fencer were consistent, while the PI case is more relevant from the practical point of view, as it is more advantageous for developing a system for recognition of the actions performed by previously unseen athletes. At the beginning we evaluated each proposed feature set separately and then we conducted a validation using the feature fusion. Then we compared the obtained results with the results achieved by state–of–the–art general action recognition methods.

6.1. Separate Feature Sets

At the beginning, we evaluated the acceleration features (Acc) using DTW and SVM with both linear and with RBF kernel. The best parameters of SVM, $c = 1$ for linear SVM and $c = 100$, $\gamma = 0.01$ for SVM with RBF kernel were determined in a grid search. The experimental results are presented in Table 1. As we can observe, DTW achieves superior classification accuracy for the PD case, although it performs poorly in the PI case. The SVM, due to its much better generalization capabilities, achieved significantly better classification accuracy in the PI case, albeit at a cost of slightly worse recognition rate in the PD case. It is worth mentioning that a high classification accuracy in the PD / DTW case proves that each fencer performed repetitions of particular actions in a consistent manner. Therefore, we can conclude that the results in the PI case are influenced mostly by the inter–person differences and not by inaccurately performed actions.

The experiments with JD and LTI feature sets were performed using SVM, both with linear and with RBF kernel. Since the non–linear kernel does not lead to statistically significant improvement of the classification accuracy, we present results for the linear SVM only. Similarly as in the case of Acc features we performed parameter tuning, which resulted in selecting the parameter $c = 1$. Results of the experiments with all feature sets (Acc, JD, LTI) using the linear SVM and RF are presented in Table 2. As we can observe, the results achieved by the SVM are better. In the PD case the classification accuracy is similar for all feature sets. In the PI case the JD features give superior results and achieve almost 80% classification accuracy. Acc features give the least accurate recognition results, probably due to being limited to input data from only one joint, i.e. the knee of the 'front' leg.

Table 2: Accuracy [%] of recognition using RF and linear SVM, with Acc, JD and LTI features on the Fencing Footwork Dataset.

	SVM		RF	
features	PD	PI	PD	PI
Acc	**93.88**	**70.71**	90.58	63.73
JD	**93.55**	**79.82**	92.07	**79.82**
LTI	94.05	**74.62**	**94.21**	72.34

In order to remove possible redundancies in feature representation, we analyzed the accuracies, which were obtained using features for all eight joints as well as for subsets of six joints, i.e. without hips, knees, ankles and feet, respectively, see also first column in Tab. 3. As we can observe, the best results were achieved without features extracted on the ankles. In the remainder of this section, we utilize features extracted on hips, knees, and feet.

Even though the JD features seem to outperform the other ones, it is worth analyzing the confusion matrices for each feature set. These are presented in Tables 4, 5 and 6. First of all, we can observe that the step actions (step forward and step backward) are relatively easily recognizable. As described in Section 3, these actions are distinctively different from the four lunge actions.

Table 3: Comparison of recognition accuracy [%] for all eight joints and subset of six joints (without hips, knees, ankles and feet, respectively).

	JD		LTI	
features	PD	PI	PD	PI
all	93.55	79.82	94.05	74.62
w/o hips	92.23	79.06	92.40	74.42
w/o knees	93.06	77.69	92.56	70.71
w/o ankles	**94.21**	**80.42**	**94.88**	**77.24**
w/o feet	93.55	79.36	93.55	74.36

The lunge actions, on the other hand, are much more difficult to recognize correctly. Particularly, the incremental speed lunge (IS) is the most difficult to recognize since in all cases the accuracy obtained for this action is close to 50%. This reason for this is due to specifics of the IS action. All lunge actions have some distinctive feature - the R action is the quickest one, the W/W action has a pause, and the JS is the longest. The IS action is characterized by a specific acceleration pattern, which is both difficult to capture and similar to other actions, as the acceleration stage is always present. An important observation is that the best recognition rates for particular actions are not present in a single feature set. Different feature sets seem to be best suited for some actions. LTI features are superior for recognition of the most difficult to recognize IS action, while JD features significantly outperform other features in the case of W/W and JS lunge actions. The Acc features, despite relatively low accuracy for lunge actions, provide best recognition rates for both SF and SB actions. This leads to a conclusion, that using multiple feature sets simultaneously may improve overall accuracy.

Table 4: Confusion matrix [%] for the person independent (PI) case using acceleration features (Acc) and linear SVM.

	R	IS	W/W	JS	SF	SB
R	**75.00**	7.41	-	17.59	-	-
IS	17.12	**45.95**	9.01	27.92	-	-
W/W	-	19.30	**59.65**	21.05	-	-
JS	15.60	23.85	10.09	**50.46**	-	-
SF	0.93	0.92	-	-	**98.15**	-
SB	1.83	-	-	-	1.83	**96.34**

Table 5: Confusion matrix [%] for the person independent (PI) case using Joint Dynamics (JD) features and linear SVM.

	R	IS	W/W	JS	SF	SB
R	**79.08**	13.99	1.95	4.98	-	-
IS	17.32	**51.05**	12.71	18.92	-	-
W/W	6.14	12.28	**77.19**	4.39	-	-
JS	-	14.68	4.59	**80.73**	-	-
SF	0.92	-	-	0.93	**98.15**	-
SB	2.75	-	-	-	0.92	**96.33**

Table 6: Confusion matrix [%] for the person independent (PI) case using Local Trace Images (LTI) features and linear SVM.

	R	IS	W/W	JS	SF	SB
R	**76.20**	18.97	3.90	0.93	-	-
IS	14.77	**58.20**	8.11	18.92	-	-
W/W	12.28	23.68	**61.40**	2.64	-	-
JS	-	19.48	5.50	**75.02**	-	-
SF	-	-	-	-	**96.30**	3.70
SB	-	-	-	-	3.67	**96.33**

6.2. Feature Fusion

Usually, late feature integration [51] is preferred over early integration for two primary reasons. First, the feature concatenation in the early integration usually leads to high dimensional data space. Second, late integration provides far better flexibility in data modeling and designing the classifiers. The late integration is especially desirable if different data sources are available. In general, in early feature fusion a larger multi–modal dataset might be necessary due to higher dimension of data space.

The fusion of the proposed feature sets was performed using a SVM [23], a RF and an MLP neural network, which was described in Section 5. We considered two combinations of the feature sets. The first one included LTI and JD feature sets, as they are both extracted from the Kinect data. In the second feature combination we included Acc features as well, in order to verify if additional modality (accelerometer) improves the classification accuracy. For the MLP neural network we used a single hidden layer and we experimented with various parameters, namely number of neurons in the hidden layer n, learning rate r and momentum m. Using the grid search we found the best parameters to be $n = 11$, $r = 0.3$, $m = 0.2$ for the LTI + JD fusion and $n = 7$, $r = 0.3$, $m = 0.4$ for the LTI + JD + Acc fusion.

Experimental results, which were performed to evaluate different fusion strategies demonstrated that the MLP outperforms both SVM– and RF–based fusion. As we can observe in Tab. 7, the MLP achieves superior results both on LTT + JD and LTI + JD + Acc features.

Table 7: Evaluation of recognition accuracy [%] of actions from the Fencing Footwork Dataset for different fusion methods.

	SVM	RF	MLP
LTI+JD	76.67	78.60	**81.49**
LTI+JD+Acc	79.21	81.03	**83.69**

For comparison, we also considered state–of–the–art methods for general action recognition on our Fencing Footwork Dataset. First we implemented and evaluated the MHI algorithm [47] as it was the starting point for our LTI descriptor. Next we considered EigenJoints [38], which employ distances between joints, and hence constitute a good reference point for our JD descriptor. The third verified method is called Local Occupancy Patterns (LOP) and is based on 3D depth histograms that are computed around joints as well as Fourier Temporal Pyramid (FTP), which uses Fourier Transform in multiple size windows [27]. It differs from our JD features in that it employs relative joint positions (rather than velocity and acceleration used in the JD features) and does not include overlapping windows. The next method used for comparison is the Histogram of Oriented 4D Normals (HON4D) [37], which describes the depth map sequences using histograms capturing distribution of the surface normal orientation in the 4D space of time, depth and spatial coordinates. The C3D pre-trained model has been used to extract the motion features, which were then classified by the linear SVM [34]. In contrast to work mentioned above, where the RGB videos were used, we fed the depth data for the three input channels of the C3D network. We also compared our method with recently proposed SkeletonNet [40]. The experimental results are presented in Table 8.

We can observe that in the PD case, all methods except EigenJoints and MHI achieve accuracies better than 90%. There is also a small improvement compared to using separate feature sets. Employing only LTI features resulted in 94.05% recognition rate, see Table 2, while fusion of the JD features increased the recognition accuracy to 94.88%, see also results in 7th row in Tab. 8. In this case, adding the Acc features had little effect.

Table 8: Accuracy [%] of recognition of actions from the Fencing Footwork Dataset using fusion of Acc, JD and LTI features, compared to state-of-the-art action recognition methods.

	PD	PI
EigenJoints [38]	35.04	29.89
MHI [47]	88.60	61.25
HON4D [37]	93.22	75.87
LOP/FTP	94.21	76.14
SkeletonNet [40]	93.12	64.36
C3D [34]	94.55	67.63
LTI + JD (ours)	94.88	81.49
LTI + JD + HON4D	95.67	80.91
LTI + JD + LOP/FTP	94.55	81.39
LTI + JD + Acc (ours)	**95.80**	**83.59**
LTI + JD + Acc + HON4D	95.77	82.67
LTI + JD + Acc + LOP/FTP	95.54	81.91

In the PI case the fusion of LTI and JD features provided a noticeable improvement, as obtained accuracy was 81.49%. Adding the Acc features further increases the recognition rate to 83.59%. We can also observe that the proposed algorithm is significantly better than the state–of–the–art methods used for comparison. It is worth noting, that in general action recognition, different dynamics of execution of an action are deliberately treated as the same action. For this reason, the deep-learning approaches evaluated in this work, which demonstrated superior performance in comparison to hand crafted-based general action recognition, were not able to capture the subtle differences of motion in recognition of dynamic actions. This indicates, that the problem of recognition of dynamics of actions is in fact considerably different from the general action recognition. Since the HON4D and LOP/FTP features provide significantly better results than other examined state-of-the-art methods, we decided to examine if combining them would lead to further improvement. We examined the following cases: LTI + JD + HON4D, LTI + JD + LOP/FTP, LTI + JD + Acc + HON4D, LTI + JD + Acc + LOP/FTP. As we can observe in Tab. 8, the combined features were not able to provide higher recognition accuracy. Confusion matrix for the PI case using fusion of all the proposed feature sets is presented in Table 9. For all actions, except the JS lunge, the obtained recognition rate is better than in the case of separate feature sets, see Tables 4, 5, 6. The reason the results for the dynamic action JS are slightly worse in comparison to results in Tab. 5 is that the Acc features do not represent well the JS in comparison to Joint Dynamics representation. In particular, the R and W/W lunge actions are recognized much better. The recognition of IS lunge action is still the most difficult, although even in this case an improvement can be observed. The step actions are recognized with none (SB) or very little error (SF).

Table 9: Confusion matrix [%] for the person independent (PI) case using fused features (LTI + JD + Acc) and linear SVM.

	R	IS	W/W	JS	SF	SB
R	**87.04**	7.41	0.93	4.62	-	-
IS	17.18	**59.36**	8.18	15.28	-	-
W/W	4.61	9.65	**80.48**	4.38	-	0.88
JS	2.75	17.53	0.92	**78.90**	-	-
SF	0.92	-	-	-	**97.30**	1.78
SB	-	-	-	-	0.92	**99.08**

7. Conclusions

In this paper we addressed the problem of dynamic action recognition in sports. We considered the recognition of basic footwork in fencing on the basis of depth maps and acceleration data. We found that action dynamics is an important issue in sports actions recognition and differs significantly from typical action recognition problems. Dynamic actions with very similar trajectories but different dynamics of motion require novel methods for proper classification. We presented the first public dataset with such actions - the Fencing Footwork Dataset. This multi–modal dataset includes data collected with both the Kinect and the x–IMU sensors. In order to properly model the motion changes in dynamic actions we proposed 3 sets of features, based on the skeleton data (Joint Dynamics), skeleton and depth data (Local Trace Images) and inertial data (Acceleration features). We evaluated all feature sets separately as well as using MLP neural network for fusion. The experimental results indicate that the fusion of different features significantly improves the recognition accuracy. On the Fencing Footwork Dataset the proposed algorithms outperform state-of-the-art methods for general action recognition. We believe that analysis of dynamics of actions is a novel and interesting direction in the area of action recognition and that the presented dataset will be useful for further research of this subject.

Acknowledgements

This work was partially supported by Polish National Science Center (NCN) under a research grant 2014/15/B/ST6/02808 and a grant 11.11.230.124. We would like to thank the fencers from Aramis Fencing School (aramis.pl) for their help in acquiring the Fencing Footwork Dataset.

References

[1] P. M. McGinnis, *Biomechanics of Sport and Exercise. Human Kinetics*. Human Kinetics, 1999.

[2] M. R. Yeadon and J. H. Challis, “The future of performance-related sports biomechanics research,” *Sports Sci.*, vol. 12, no. 1, pp. 3–32, 1994.

[3] D. V. Knudson, *Qualitative diagnosis of human movement: improving performance in sport and exercise*. Human Kinetics, 2013.

[4] J. Margarito, R. Helaoui, A. M. Bianchi, F. Sartor, and A. G. Bonomi, “User-independent recognition of sports activities from a single wrist-worn accelerometer: A template-matching-based approach,” *IEEE Trans. on Biomedical Engineering*, vol. 63, no. 4, pp. 788–796, 2016.

[5] M.-S. Pan, K.-C. Huang, T.-H. Lu, and Z.-Y. Lin, “Using accelerometer for counting and identifying swimming strokes,” *Pervasive and Mobile Computing*, vol. 31, no. Supp. C, pp. 37 – 49, 2016.

[6] F. Zhu, L. Shao, J. Xie, and Y. Fang, “From handcrafted to learned representations for human action recognition,” *Image Vision Comput.*, vol. 55, no. P2, pp. 42–52, 2016.

[7] D. Weinland, R. Ronfard, and E. Boyer, “A survey of vision-based methods for action representation, segmentation and recognition,” *Comput. Vis. Image Underst.*, vol. 115, no. 2, pp. 224–241, 2011.

[8] K. Soomro and A. R. Zamir, *Computer Vision in Sports*. Cham: Springer Int. Publ., 2014, ch. Action Recognition in Realistic Sports Videos, pp. 181–208.

[9] W. Li, Z. Zhang, and Z. Liu, “Action recognition based on a bag of 3D points,” in *IEEE Computer Society Conf. on Computer Vision and Pattern Recognition - Workshops*, June 2010, pp. 9–14.

[10] T. Mauthner, C. Koch, M. Tilp, and H. Bischof, “Visual tracking of athletes in beach volleyball using a single camera,” *Int. J. of Computer Science in Sport*, vol. 6, no. 2, pp. 21–34, 10 2008.

[11] P. Hämäläinen, “Interactive video mirrors for sports training,” in *Proc. of the Third Nordic Conf. on Human-computer Interaction*. New York, NY, USA: ACM, 2004, pp. 199–202.

[12] R. Urtasun, D. J. Fleet, and P. Fua, “Monocular 3D tracking of the golf swing,” in *IEEE Comp. Society Conf. on Computer Vision and Pattern Recognition*, 2005, pp. 932–938 vol. 2.

[13] T. Hachaj, M. R. Ogiela, and K. Koptyra, “Effectiveness comparison of Kinect and Kinect 2 for recognition of Oyama karate techniques,” in *Proc. of 18th Int. Conf. on Network-Based Information Systems*. Washington, DC, USA: IEEE Computer Society, 2015, pp. 332–337.

[14] L. Zhang, J. C. Hsieh, T. T. Ting, Y. C. Huang, Y. C. Ho, and L. K. Ku, “A Kinect based golf swing score and grade system using GMM and SVM,” in *5th Int. Congress on Image and Signal Processing*, 2012, pp. 711–715.

[15] S. Gourgari, G. Goudelis, K. Karpouzis, and S. Kollias, “Thetis: Three dimensional tennis shots a human action dataset,” in *IEEE Conf. on Computer Vision and Pattern Recognition Workshops*, 2013, pp. 676–681.

[16] L. Lo Presti and M. La Cascia, “3D skeleton-based human action classification,” *Pattern Recogn.*, vol. 53, pp. 130–147, 2016.

[17] L. Lo Presti, M. La Cascia, S. Sclaroff, and O. Camps, "Hankelet-based dynamical systems modeling for 3D action recognition," *Image Vision Comput.*, vol. 44, pp. 29–43, 2015.

[18] S. Noiumkar and S. Tirakoat, "Use of optical motion capture in sports science: A case study of golf swing," in *Int. Conf. on Informatics and Creative Multimedia*, 2013, pp. 310–313.

[19] A. Avci, S. Bosch, M. Marin-Perianu, R. Marin-Perianu, and P. Havinga, "Activity recognition using inertial sensing for healthcare, wellbeing and sports applications: A survey," in *23th Int. Conf. on Architecture of Computing Systems*, 2010, pp. 1–10.

[20] K. King, S. Yoon, N. Perkins, and K. Najafi, "Wireless MEMS inertial sensor system for golf swing dynamics," *Sensors and Actuators A: Physical*, vol. 141, no. 2, pp. 619 – 630, 2008.

[21] F. Eckardt, A. Mnz, and K. Witte, "Application of a full body inertial measurement system in dressage riding," *J. of Equine Veterinary Science*, vol. 34, no. 11, pp. 1294 – 1299, 2014.

[22] C. N. Nam, H. J. Kang, and Y. S. Suh, "Golf swing motion tracking using inertial sensors and a stereo camera," *IEEE Trans. on Instrumentation and Measurement*, vol. 63, no. 4, pp. 943–952, 2014.

[23] M. S. Cheema, A. Eweiwi, and C. Bauckhage, *A Stochastic Late Fusion Approach to Human Action Recognition in Unconstrained Images and Videos*. Springer, 2014, pp. 616–628.

[24] V. Bloom, D. Makris, and V. Argyriou, "G3D: A gaming action dataset and real time action recognition evaluation framework," in *2012 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition Workshops*, 2012, pp. 7–12. [Online]. Available: http://dx.doi.org/10.1109/CVPRW.2012.6239175

[25] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: A local SVM approach," in *Proc. Int. Conf. on Pattern Recognition, vol. 3*. Washington, DC, USA: IEEE Computer Society, 2004, pp. 32–36.

[26] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 12, pp. 2247–2253, 2007.

[27] Y. Wu, "Mining actionlet ensemble for action recognition with depth cameras," in *IEEE Conf. on Computer Vision and Pattern Recognition*. Washington, DC, USA: IEEE Computer Society, 2012, pp. 1290–1297.

[28] Z. Borysiuk, K. Piechota, and T. Minkiewicz, "Analysis of performance of the fencing lunge with regard to the difficulty level of a technical-tactical task," *J. of Combat Sports and Martial Arts*, vol. 4, pp. 135–139, 12 2013.

[29] K. C. Moore, F. M. Chow, and J. Y. Chow, "Novel lunge biomechanics in modern sabre fencing," *Proc. Eng.*, vol. 112, pp. 473 – 478, 2015, 'The Impact of Technology on Sport VI' 7th Asia-Pacific Congress on Sports Technology.

[30] M. Gholipour, A. Tabrizi, , and F. Farahmand, "Kinematics analysis of lunge fencing using stereophotogrametry," *Sports Sci.*, vol. 2, no. 1, pp. 32–37, 2008.

[31] G. Mantovani, A. Ravaschio, P. Piaggi, and A. Landi, "Fine classification of complex motion pattern in fencing," *Proc. Eng.*, vol. 2, no. 2, pp. 3423–3428, 2010.

[32] J. M. Chaquet, E. J. Carmona, and A. Fernndez-Caballero, "A survey of video datasets for human action and activity recognition," *Computer Vision and Image Understanding*, vol. 117, no. 6, pp. 633 – 659, 2013.

[33] Y. G. Jiang, Q. Dai, W. Liu, X. Xue, and C. W. Ngo, "Human action recognition in unconstrained videos by explicit motion modeling," vol. 24, no. 11, pp. 3781–3795, 2015.

[34] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *2015 IEEE Int. Conf. on Computer Vision (ICCV)*, 2015, pp. 4489–4497. [Online]. Available: http://dx.doi.org/10.1109/ICCV.2015.510

[35] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *2014 IEEE Conf. on Computer Vision and Pattern Recognition*, 2014, pp. 1725–1732. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2014.223

[36] L. Chen, H. Wei, and J. Ferryman, "A survey of human motion analysis using depth imagery," *Pattern Recogn. Lett.*, vol. 34, no. 15, pp. 1995–2006, 2013.

[37] O. Oreifej and Z. Liu, "HON4D: Histogram of oriented 4D normals for activity recognition from depth sequences," in *IEEE Conf. on Computer Vision and Pattern Recognition*. Washington, DC, USA: IEEE Computer Society, 2013, pp. 716–723.

[38] X. Yang and Y. Tian, "Effective 3D action recognition using EigenJoints," *J. of Visual Communication and Image Representation*, vol. 25, no. 1, pp. 2 – 11, 2014.

[39] S. Herath, M. Harandi, and F. Porikli, "Going deeper into action recognition: A survey," *Image and Vision Computing*, vol. 60, pp. 4 – 21, 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0262885617300343

[40] Q. Ke, S. An, M. Bennamoun, F. A. Sohel, and F. Boussaïd, "SkeletonNet: Mining deep part features for 3-D action recognition," *IEEE Signal Process. Lett.*, vol. 24, no. 6, pp. 731–735, 2017.

[41] Z. Czajkowski, *Understanding fencing: The unity of theory and practice*. SKA Swordplay Books, 2005.

[42] The x-IMU Inertial Measurement Unit. [Online]. Available: http://www.x-io.co.uk/products/x-imu/

[43] O. D. Lara and M. A. Labrador, "A survey on human activity recognition using wearable sensors," *IEEE Communications Surveys and Tutorials*, vol. 15, pp. 1192–1209, 2013.

[44] R. Zhu and Z. Zhou, "A real-time articulated human motion tracking using tri-axis inertial/magnetic sensor package," *IEEE Trans. on Neural Systems and Rehabilitation Engineering*, vol. 12, pp. 295–302, 2004.

[45] F. Malawski and B. Kwolek, "Classification of basic footwork in fencing using accelerometer," in *Int. Conf. on Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA)*, 2015, pp. 51–55.

[46] J. P. Amaro and P. Sergio, "A survey of sensor fusion algorithms for sport and health monitoring applications," in *42nd Annu. Conf. IEEE Ind. Electron. Soc.*, 2016, pp. 5171–5176.

[47] A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 3, pp. 257–267, 2001.

[48] L. Shao, L. Ji, Y. Liu, and J. Zhang, "Human action segmentation and recognition via motion and shape analysis," *Pattern Recogn. Lett.*, vol. 33, no. 4, pp. 438–445, 2012.

[49] D. J. Berndt and J. Clifford, "Using dynamic time warping to find patterns in time series," in *Proc. of the 3rd Int. Conf. on Knowledge Discovery and Data Mining*. AAAI Press, 1994, pp. 359–370.

[50] J. Platt, "Probabilistic outputs for Support Vector Machines and comparison to regularized likelihood methods," in *Advances in Large Margin Classifiers*, vol. 10, no. 3, 2000, pp. 61–74.

[51] L. I. Kuncheva, "A theoretical study on six classifier fusion strategies," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 2, pp. 281–286, 2002.