Improving Multimodal Action Representation with Joint Motion History Context

Filip Malawski[a,*], Bogdan Kwolek[a]

[a] *AGH University of Science and Technology, 30 Mickiewicza Av., 30-059 Krakow, Poland*

Abstract

Automatic recognition of actions can be addressed by employing data from multiple sensors, such as RGB cameras, depth sensors or inertial measurement units. Recent studies show that multimodal representations of actions are effective in providing rich information about motion patterns. In this work, we propose a novel action descriptor, called Joint Motion History Context, which is based on depth and skeleton data. It improves action representation when used with previously introduced descriptors that are based on depth, skeleton and inertial data. A feature selection method is proposed as well, which ranks features on the basis of their inter-class discriminative power, while minimizing redundancy in the selected feature subset. Decision-level fusion, based on Support Vector Machines and Multilayer Perceptron is employed to effectively combine motion pattern information from multiple feature sets. Experimental results on two publicly available datasets, FFD and UTD-MHAD, demonstrated that the proposed methods outperform state-of-the-art algorithms.

Keywords: Action recognition, Action descriptors, Depth maps, Feature selection, Multimodal representation, Decision-level fusion

*Correspondence to: Department of Computer Science, AGH University of Science and Technology, 30 Mickiewicza Av., Building D-17, 30-059 Krakow, Poland

Email address: fmal@agh.edu.pl (Filip Malawski)

1. Introduction

In recent years much effort has been put into development of methods dedicated to automatic action recognition [1]. Providing machines with the ability to understand human motion yields many possibilities. Video games industry has already adapted action recognition for creating immersive experience for players, mainly with the use of the Kinect sensor [2]. This device has been also employed in health related scenarios, such as exercise assistance for elderly people [3] or rehabilitation monitoring [4]. Automatic surveillance and abnormal activity detection is another possible application area[5, 6]. Human-machine interfaces employ action recognition in smart homes [7] and in robots dedicated to assist humans in various tasks [8]. Another interesting field for applying automatic human motion analysis is sport [9]. Valuable feedback for players and coaches can be provided with action recognition [10], as well as qualitative analysis [11, 12].

Initial approaches to action recognition employed solely RGB cameras [13]. However, effectiveness of such methods is limited due to vulnerability to changing lighting conditions, as well as difficulties related to obtaining reliable 3D motion information. With launch of the Kinect sensor, the use of depth maps and skeletal representation of the human motion became popular [14]. Depth maps provide 3D information that is obtained with an active infra-red sensor, and therefore are robust to changing light. On the other hand, their range and resolution are more limited than RGB cameras. They are also prone to errors related to external infra-red sources, including sunlight [15]. Another approach to action recognition is to employ inertial measurement units (IMU) [16]. While extracting exact position is not easy, mainly due to the accumulation of error during integration, IMUs provide precise acceleration and rotation data, both of which can be successfully applied for action recognition [17]. Moreover, IMUs usually deliver data at higher sampling frequency than visual sensors. On the other hand, they need to be mounted on the subject, which might be a considerable limitation in some scenarios.

Information provided by each of the aforementioned modalities - RGB, depth, skeleton and inertial, is very diverse. Therefore, it is often beneficial to combine those data in order to increase effectiveness of action recognition process [18]. Methods operating on multimodal data have been shown to be efficient in many scenarios [19, 20]. Another effective manner of addressing action recognition is combining multiple different features extracted from the same modality. Such an approach to action recognition has been proposed for the RGB data [21, 22].

In this work we employ fusion of multiple different features in a multimodal scenario. We first propose Joint Motion Context History descriptor (JMHC), which is computed on the basis of depth maps and skeleton data. JMHC complements other, previously proposed descriptors, namely Joint Dynamics (JD) and Local Trace Images (LTI), both of which are based on skeleton data, as well as Acc, which is based on accelerometric data [20]. RGB data is not considered, as it might not be reliable enough in challenging scenarios, for instance in sports, where poor lighting and cluttered background considerably hinder person extraction [20]. Our novel descriptor generates high-dimensional feature vector, therefore effective feature selection method is proposed as well. Decision-level fusion is applied, by using a separate Support Vector Machine (SVM) for each feature set and then employing Multilayer Perceptron (MLP) for final classification. Figure 1 depicts the framework of our approach. The proposed methods are evaluated on two publicly available datasets, FFD [20] and UTD-MHAD [23]. Experimental results show that our methods outperform the state-of-the-art algorithms.

2. Related work

Considering depth-based feature extraction and action representation, the action recognition approaches can be divided into two major categories: skeleton-based and depth map-based approaches [24]. The skeleton-based approaches use high-level skeleton representation, that is extracted from depth map sequences. Although skeleton-based representations are robust for scenarios in which lit-

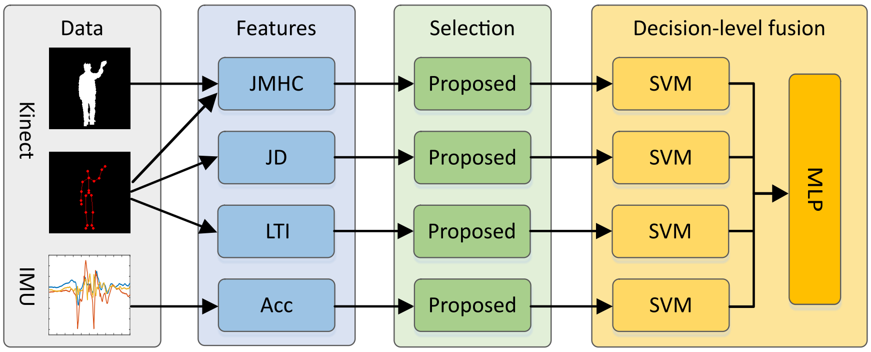

Figure 1: Framework of the proposed approach.

tle occlusion occurs, they can produce inaccurate results or even they may fail when self-occlusions or occlusions occur. What is more, skeletal joint data are not always available, particularly in scenarios, where a ceiling-mounted or a pan-tilt depth sensor is utilized to observe larger areas [25]. Thus, a lot of efforts have been undertaken to develop features for robust action recognition in depth maps.

In [26] a Sequence of Most Informative Joints (SMIJ) representation is calculated on the basis of joint angle time series. Joints with the most discriminatory power are found in temporal windows by using measures like mean or variance of joint angles, or angular velocity of joints. Based on that, the joints are rank-ordered with regard to temporal segments and every action sequence is encoded as the set of N most informative joints in each window. In [27] authors attempt to minimize the number of frames required to recognize actions. Current and previous frames, as well as a general pose representing a person at rest, are employed in order to compute feature vector consisting of Euclidean distances between joint pairs. Each feature is then clustered into one of five groups via the k-means algorithm. Finally, logistic regression is used to automatically determine distinctive canonical poses for each action. Authors of [28] also utilize 3D position differences of the joint pairs: in the current frame, between the current frame and the former frame, between the current frame and the initial one. Afterwards, the Principal Component Analysis (PCA) is applied to the concatenated feature vector in order to extract EigenJoints, which encode the human pose information for recognition of actions. In [29] key joints are selected based on their information gain, then weighted position and velocity histograms are composed with trajectory features and the final descriptor is encoded with Fisher Vectors (FV). Deep learning techniques are employed for action recognition using skeleton-based features as well. The authors of [30] capture spatial structure of the body by computing cosine distance and normalized magnitude for vectors describing joint positions with respect to base joints in five body parts, and then employ a Convolutional Neural Network (CNN) for classification. In [31] a skeleton-based clip representation of actions is proposed, which captures both spatial and temporal information. Based on that representation, classification is performed with a Multitask Convolutional Neural Network (MTCNN). The authors of [32] evaluate several neural networks for the task of action prediction and propose a novel windows scale selection scheme, which allows the network to focus on the performed part of the ongoing action.

Depth map-based space-time features are extracted by considering each action sequence as a 4D volume along spatial (x, y, z) and temporal t coordinates. The depth map sequence can be represented either as a whole or as a set of local feature descriptors. Such methods showed promising results, nevertheless, subtle motions might not achieve satisfactory recognition performance. In [33], depth video sequences are described by a histogram that captures the distribution of the surface normal orientation in the 4D space with time, depth and spatial coordinates. In order to build the histogram of oriented 4D surface normals (HON4D), 4D projectors are created, which quantize the 4D space and represent the possible directions for the 4D normal. In [34], Depth Motion Maps (DMM) were proposed to capture the aggregated temporal motion energies. The 3D silhouettes are projected onto three pre-defined orthogonal planes and then normalized. The DMM-HOG descriptors are constructed by concatenating the Histogram of Oriented Gradients (HOG) features from summed binary maps on each plane. In [35], a representation for 3D action recognition, called Space-Time Occupancy Patterns (STOP) has been proposed. It describes the 4D space-time patterns of human actions by partitioning the 4D video volume into 4D space-time cells and then aggregating the occupancy information in each cell. A deep CNN is employed in [36] to directly learn spatio-temporal features from raw depth sequences. The depth videos are preprocessed by normalization to cuboids of fixed size. For final classification, SVM and late fusion of skeleton features are used.

Employing inertial and magnetic data from body-worn sensors is another viable approach to action recognition. In [37], 9-axis IMU sensor is evaluated for the purpose of gait recognition, by comparison with results obtained on the basis of stereophotogrammetry. Experiments show, that such sensor is a valid tool for gait analysis. A popular approach to action recognition from inertial data includes division to time windows and extraction of features in time or frequency domain [38]. Time domain features include statistical measures, such as mean, variance, etc. [39]. Transformation to frequency domain is usually performed by applying Discrete Cosine Transform (DCT) or Fast Fourier Transform (FFT) [40]. Wavelets can be employed as well [41]. IMU sensors are particularly useful in automatic sport analysis, and have been employed for instance for recognition of swimming strokes [42] or fencing footwork [43].

Fusion of depth and inertial data have been explored by researchers in order to improve action recognition accuracy [44]. In [19] DMM features for three orthogonal planes are extracted from depth map sequences, whereas statistical measures are computed in time windows from accelerometric data. Both feature-level and decision-level fusion strategies with multiple classifiers were investigated. Better recognition accuracy is achieved when employing both modalities in comparison to using each one separately. Indoor activity recognition is addressed in [45]. Authors employ five three-axis accelerometers as well as joint positions provided by the Kinect. Acquired data is divided into temporal windows and then classified by an ensemble of binary one-vs-all neural network classifiers. An approach proposed in [46] employs both depth and inertial data for reliable fall detection. Data from body-worn accelerometer are used to detect abnormal situations and depth data is used to verify such detections in order to minimize the number of false alarms.

In this work we consider fusion of depth, skeleton and inertial-based modalities, by employing decision-level fusion not only for combining features obtained from different modalities, but also for combining different types of features obtained from the same modality. We employ action descriptors that were introduced in our previous work [20] and propose a novel, depth map and skeleton-based descriptor, which complements the previous ones and improves recognition accuracy. The proposed methods are evaluated on two publicly available multimodal datasets, FFD [20] and UTD-MHAD [23], both of which include depth, skeletal and inertial data.

3. Proposed descriptors

In this section we describe the proposed Joint Motion Context descriptor designed to model human poses on a single frame and then we extend it to Joint Motion History Context descriptor, which incorporates more temporal information. Final feature vector describing an entire action sequence is computed by using time window-based approach.

3.1. Joint Motion Context

We propose a novel descriptor, that is called Joint Motion Context (JMC) and which employs both depth map and skeleton data simultaneously. The proposed descriptor is responsible for describing local changes in motion around the selected joints. It involves calculating a histogram-based motion descriptor spatially at every frame in box regions around joints, that are given by the skeleton tracker. Such descriptor should be more sensitive to motion differences in nearby locations to joint and therefore we employ log-polar histograms. The log-polar histograms were used, among others, in the well known Shape Context to describe static features of image edges [47].

In our approach we employ the depth data provided by the Kinect sensor. We utilize silhouette of a person, extracted automatically by the Kinect SDK, as well as the skeleton data. The silhouettes, represented as binary images, are used to compute motion difference between two consecutive frames, see Fig. 2. As demonstrated in [48], the difference between the intensity images can be very useful for action representation as it does not require detection/location of bounding boxes. In this work the inter-frame motions are determined on the basis of consecutive depth maps.

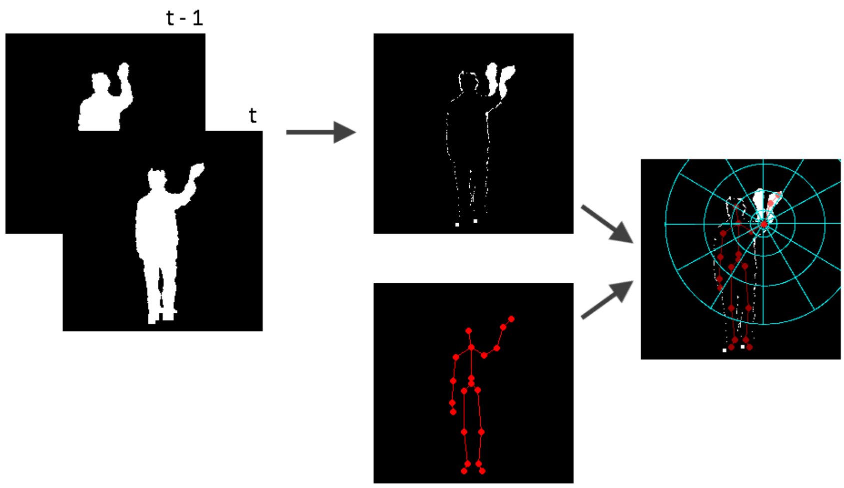

Figure 2: JMC descriptor. Motion histograms for each joint are computed on the basis of differences of silhouettes from consecutive frames, as well as skeleton data.

Given N points $p_1, p_2, \ldots, p_N$ representing silhouette changes between two consecutive depth maps, the motion context at joint q_i is defined as a log-polar histogram h_i of the relative points, that indicate the silhouette motion. Given l - log coordinate and φ - polar coordinate of the histogram:

$$h_i(l, \varphi) = \#\{p_j \,|\, i \neq j, (log(p_j - q_i), angle(p_j, q_i)) \in bin_{l,\varphi}\}. \tag{1}$$

This means that relative coordinates of each motion point with regard to the joint location are binned into a log-polar histogram, see Fig. 2. The histograms are then normalized. Given N_i - total number of points in histogram h_i, N_b - number of bins in a single histogram, b_k - bins, normalization is performed as follows:

$$\forall k \in 1, ..., N_b, b_k = b_k / N_i \tag{2}$$

Therefore each histogram forms a probability distribution:

$$\sum_{k=1}^{N_b} b_k = 1 \tag{3}$$

Histograms computed for each joint are then concatenated to form a single descriptor.

As demonstrated in [34] the orthogonal projections of depth data can provide useful information. On the basis of depth data it is relatively easy to create two additional, orthogonal projections of silhouettes (side-view and top-view, see Fig. 3) and hence to extend our method to include 3D motion information. We consider each of the three planes (front, side, top) separately, by computing the log-polar histograms for each joint in each plane. Therefore, the final descriptor includes complete 3D motion information.

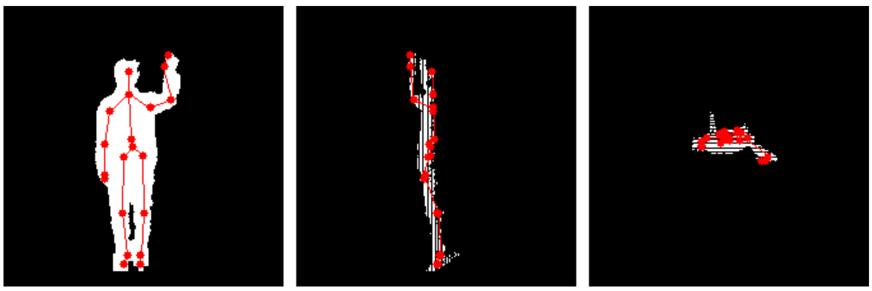

Figure 3: Silhouette and skeleton data projected onto 3 orthogonal planes. From left: front view, side view and top view.

The number of divisions is 12 in the polar coordinate, and 5 in the log coordinate. This results in total of 60 bins per descriptor of a single joint motion. The Kinect provides 20 joints, which results in the descriptor size equal to 1200 features for each of the 3 separate planes, per frame.

3.2. Joint Motion History Context

The proposed Joint Motion Context descriptor utilizes motion difference between two consecutive depth maps. Although it provides relevant information to represent joint movement in action, we found that it is useful to consider also motion arising between the current and earlier frames. Experimental results demonstrated, that the use of the silhouette differences between current and several preceding frames yields better recognition accuracy. Therefore, we combine multiple JMC descriptors, which are determined for multiple preceding frames. The proposed Joint Motion History Context (JMHC) descriptor is calculated as a weighted sum of the histograms corresponding to motion in several preceding frames, see Fig. 4. Based on the experimental results, we use 3 preceding frames, with weights $w_{t-1} = 0.25$, $w_{t-2} = 0.5$ and $w_{t-3} = 0.25$. The final JMHC descriptor has the same number of features as the JMC descriptor, but each histogram bin expresses the sum of points from difference images between current frame t and preceding frames t - 1, t - 2 and t - 3.

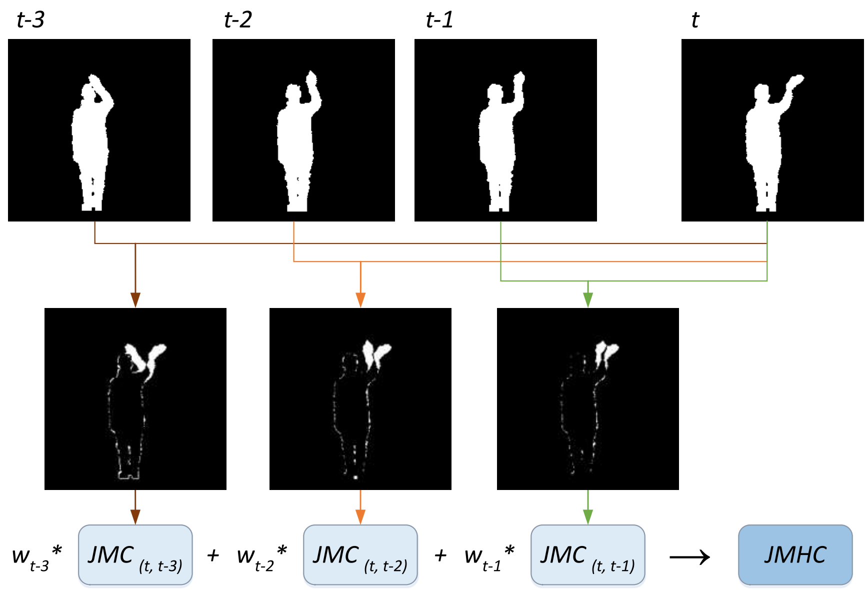

Figure 4: JMHC descriptor is constructed as a weighted sum of JMC descriptors computed using the current and 3 preceding frames.

3.3. Action Descriptor

In order to calculate informative feature descriptor for an entire action we first compute JMHC descriptor in each frame and then extract higher level features for the entire sequence. The number of frames in the considered actions may vary, therefore we first interpolate the sequence of per-frame features to a common length. Since the datasets that were employed in the experiments were recorded with the Kinect operating at 30 Hz and most actions last about 2 seconds, we interpolate every action to the length of 64 frames. The interpolation is performed on JMHC feature vectors computed for each frame. In the next step we divide the interpolated feature vector to equal-size time intervals with 50% overlap between adjacent segments, see Fig. 5. This technique has been previously used in [40], with and without overlapping. Splitting feature vectors into time windows gives the possibility to model the actions with relationship to time. In the discussed approach, segments of size 16 are used, with 50% overlapping, which results in 7 windows per data sequence. The length of the employed segments was selected experimentally.

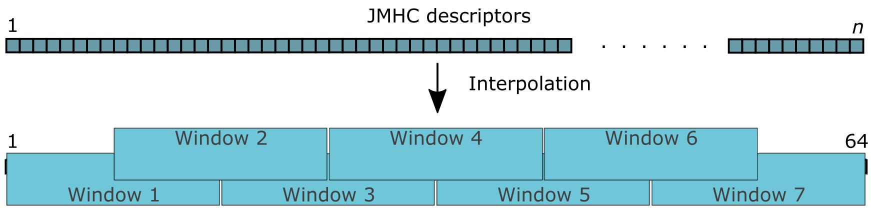

Figure 5: Action descriptor computed for a sequence consisting of n frames. Each square in the top row represents a JMHC descriptor of a single frame. The total feature vector is interpolated to a common length and then divided to equal-size overlapping windows, in which statistical descriptors are computed.

For each frame, in each window, we compute statistical measurements, namely mean and root mean square (RMS) values. Therefore, in each window the number of features equals twice the number of features for a single frame. Using JMHC descriptor the total length of feature vector describing a single action is equal to 16800 features for each plane. It is worth noting, that many of these features are expected to be irrelevant, as many bins of the log-polar histograms cover areas where no motion difference is present. Therefore, a feature selection has been employed to determine the most relevant and informative feature subset.

4. Classification

The proposed action descriptor produces a large set of features, which should be reduced before classification in order to simplify the learned models, avoid the curse of dimensionality and particularly to obtain better generalization by reducing overfitting [49]. After evaluating popular approaches for feature selection (AdaBoost [50, 51], Lasso [52, 51]) and dimensionality reduction (PCA [53]), that gave unsatisfactory results, we decided to propose a novel feature selection method, based on feature correlation. The proposed algorithm is described in Section 4.1. Action classification is performed by training separate SVM models for each employed feature set, and then applying decision-level fusion by using MLP. Details are presented in Section 4.2.

4.1. Feature selection

In pattern recognition a feature selection is frequently used to identify and remove unneeded, irrelevant and redundant attributes from data that do not contribute to the classification performance of the predictive model. Wrapper feature selection methods train and evaluate a model for a large number of subsets, which is computationally expensive [49]. Filter methods, on the other hand, rank features based on a selected metric, which is much faster. Our algorithm is based on the idea of scoring class-dependent features [54]. The proposed method ranks features on the basis of a distance between histograms computed for each feature per each class. It also reduces redundancy in the selected subset, by considering, in each iteration, correlation to the already selected features. The main steps of the algorithm are as follows:

Step 1:

Create matrix H of normalized histograms per each feature and class. Each histogram is a probability distribution of a given feature in a given class.

C - set of all classes

F - set of all features

$h_{i,j}$ - histogram for $i-th$ feature and $j-th$ class

H:

	1	2	...	$size(C)$
1	$h_{1,1}$	$h_{1,2}$	...	$h_{1,size(C)}$
2	$h_{2,1}$	$h_{2,2}$	...	...
...	...	...	...	...
$size(F)$	$h_{size(F),1}$	$h_{size(F),2}$	...	$h_{size(F),size(C)}$

where

$size(H) = size(F) * size(C)$

Step 2:

Compute matrix A of weights a describing how well each feature is able to distinguish each pair of classes.

P - set of all distinctive pairs of classes

$P = \{(c_i, c_j) \mid i < j, i = 1...size(C), j = 1...size(C)\}$

$a_{i,k}$ - weight for $i-th$ feature and $k-th$ pair of classes

d - distance metric between two histograms h_1 and h_2, with N_b bins each $d(h_1, h_2) = \sum_{i=1}^{N_b} abs(h_1(i) - h_2(i))$

for i = 1:size(F)

    for k = 1:size(P)

        c_1 = *first class of k-th pair*

        c_2 = *second class of k-th pair*

        $a_{i,k} = d(h_{i,c_1}, h_{i,c_2})$

where:

$size(P) = \frac{size(C)*(size(C)-1)}{2}$

$size(A) = size(F) * size(P)$

Step 3:

Initialize vector B of weights b describing how much each feature is correlated to the already selected features. Initially no features are selected, therefore all weights are set to zero.

for i = *1:size*(F)
  b_i = *0*

where:

$size(B) = size(F)$

Step 4:

Iteratively select m features, by going in loop over all pairs of classes and choosing most discriminative features based on sum of weights A and B. In each step weights B are updated by adding sums of distances between each remaining features and the last selected feature. Moreover, weights B are normalized according to the current time step.

F_S - set of already selected features

F_R - set of remaining features, from which the selection is performed

t - time step

f - feature selected in the current time step

$F_S = \emptyset$

$F_R = F$

$t = 0$

while $size(F_S) < m$

    for k = *1:size*(P)

        $f = \underset{i \in F_R}{argmax}(a_{i,k} + b_i)$

        $F_S = F_S \cup f$

        $F_R = F_R \setminus f$

        for i = *1:size*(F_R)

            $b_i = \frac{t}{t+1} * b_i + \frac{1}{t+1} * \sum_{c=1}^{size(C)} d(h_{i,c}, h_{f,c})$

        $t = t + 1$

4.2. Fusion

In order to improve the action classification accuracy, decision-level fusion is applied, based on the method described in [20]. In the first step, multiple feature sets are computed, using different descriptors and modalities. For each feature set the selection is performed and a separate SVM model is learned and then used for classification. A multi-class SVM is employed to provide probabilities for each class. These are fed to MLP, which performs decision-level fusion. Fusion process is illustrated in Fig. 6.

Apart from the proposed JMHC descriptor, we employ a number of other feature extraction methods, described in our previous work [20]. Those include: Local Trace Images (LTI), Joint Dynamics (JD) and accelerometric features (Acc). LTI features describe motion of each considered joint in the estimated skeleton by creating probabilistic images of motion patterns. Positions of joints are modelled by two-dimensional normal distribution. Superposition of such Gaussians forms the image with motion pattern. JD features are computed using acceleration and velocity of the joints in the skeleton data. Short Time Fourier

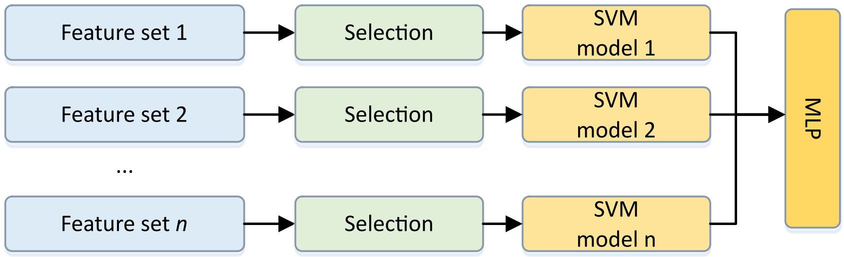

Figure 6: Fusion of n feature sets. For each feature set selection is applied and then a separate SVM model is created. MLP is trained on SVMs outputs.

Transform (STFT) is employed in overlapping temporal windows, on 3 different levels, each with different length of windows. First 3 STFT coefficients from each time segment are included in the final feature vector. Accelerometric data is preprocessed by applying high-pass filter and computing derivative of the signal. Acc features are extracted in overlapping time windows, similarly to JD features, but instead of employing frequency domain, statistical measurements in time domain are calculated - mean and RMS values for each axis and magnitude. Original LTI and JD features were computed in 2D, due to the specifics of dataset for which they were designed. However, in this work, for experiments on the UTD-MHAD dataset we extend them to 3D by computing separate feature vectors for three orthogonal planes.

5. Experimental results

Experiments were performed on two publicly available datasets, namely Fencing Footwork Dataset (FFD) [20] and UTD-MHAD [23]. Since our goal was to improve action recognition by feature fusion, we focused on datasets containing multimodal data. Both FFD and UTD-MHAD include depth, skeleton and inertial data, and the latter also RGB data. Experiments were conducted accordingly to the evaluation schemes recommended by the authors of each dataset. Weka implementation of the SVM and MLP classifiers was employed [55] in all evaluations presented below.

5.1. Datasets

FFD contains basic actions from footwork training routine in fencing, namely steps and lunges, see Fig. 7. There are six actions in the dataset: *rapid lunge (R), incremental speed lunge (IS), lunge with waiting(W/W), jumping-sliding lunge (JS), step forward (SF), step backward (SB).* Four types of lunge action (R, IS, W/W, JS) are very similar to each other and need to be distinguished by dynamics rather than trajectories, hence making this dataset challenging. Recordings were made with 10 persons performing 10-11 repetitions of each action. Depth, skeleton and inertial data were acquired by the Kinect and 9-axis x-IMU sensor respectively (RGB data is not available). In previous work we presented results for both person dependent and person independent scenarios, whereas in this work we focus only on the person independent scenario, as it is more useful in practice and challenging. The proposed algorithm has been evaluated using 10-fold leave-one-out cross-validation, where in each fold one person is used for testing and the other nine persons are used for training.

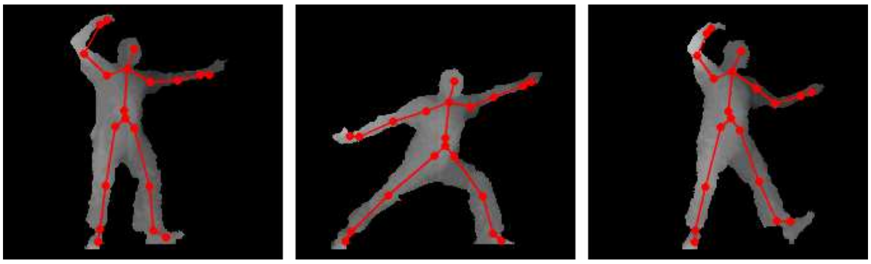

Figure 7: Sample actions from FFD - depth and skeleton view. From left: fencing stance (starting pose in all actions), lunge (pose common for all lunge types), step forward.

UTD-MHAD includes 27 general actions, with one-hand, two-hand or leg motion. Those are: *swipe left, swipe right, wave, clap, throw, arm cross, basketball shoot, draw X, draw circle clockwise, draw circle counter clockwise, draw triangle, bowling, boxing, baseball swing, tennis swing, arm curl, tennis serve, push, knock, catch, pickup throw, jog, walk, sit to stand, stand to sit, lunge,* *squat*, see Fig. 8. The actions were performed by 8 persons, with 4 repetitions. The dataset includes RGB, depth and skeleton data acquired by the Kinect, as well as inertial data acquired with a low-cost inertial sensor. The evaluation protocol recommends using subjects number 1, 3, 5, 7 for training and subjects number 2, 4, 6, 8 for testing.

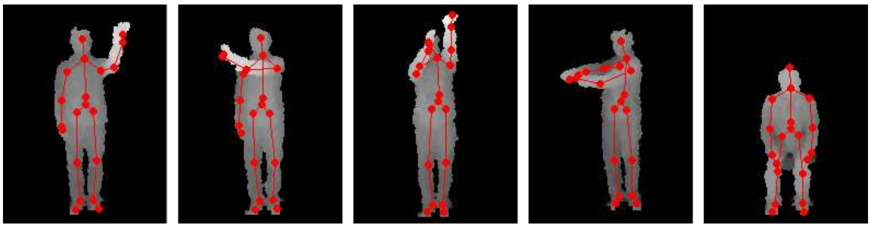

Figure 8: Sample actions from UTD-MHAD - depth and skeleton view. From left: wave, swipe left, basketball shoot, baseball swing, sit to stand.

5.2. FFD results

Features for the FFD dataset were computed in 2D and only for the lower body (feet, ankles, knees, hips), as proposed in the original paper introducing the dataset [20]. This is due to the specifics of the fencing footwork actions, for which motion in the other planes is not relevant and motion of the upper body may vary for the same footwork action. Person independent scenario is evaluated, as the more challenging one. All considered feature sets (LTI, JD, Acc, JMHC) are used in data fusion.

At first, the proposed feature selection algorithm was evaluated for individual feature sets using linear SVM classifier, see Table 1. For the SVM, the parameter $C = 1$ was employed, as the result obtained in grid search for the best parameter. PCA, AdaBoost and Lasso algorithms were used to evaluate and compare the performances of dimensionality reduction and feature selection. For comparison, results without feature selection are also included. The presented results contain recognition accuracy, as well as the number of features selected by each method. The proposed method outperformed the other ones on JD and JMHC feature sets. In the case of Acc features, none of the methods was able to improve the accuracy, most likely for the reason that these features were compact to start with (initial size 168). For the LTI feature set, the Lasso algorithm achieved the best results.

Table 1: Recognition accuracy (%) for the FFD dataset using linear SVM ($C = 1$) and different feature selection methods on different feature sets.

		w/o sel.	PCA	AdaBoost	Lasso	Proposed
LTI	#feat	1536	100	600	500	800
	acc.	77.24	74.16	76.9	**79.03**	77.36
JD	#feat	792	200	400	600	120
	acc.	80.42	78.12	79.18	80.24	**82.52**
Acc	#feat	168	60	150	160	140
	acc.	**70.71**	68.24	70.06	70.36	70.67
JMHC	#feat	6720	100	750	650	720
	acc.	74.16	75.99	73.1	76.44	**79.03**

The proposed feature selection method has then been evaluated in the proposed fusion scenario, see Section 4.2. Feature selection was performed for each feature set separately, before training individual SVM models. The obtained results are presented in Table 2 and include comparison with other feature selection algorithms. Two scenarios were evaluated, where the first one includes only features obtained from the Kinect sensor, and the second on comprises additional inertial-based Acc features. The proposed feature selection method outperforms all other methods in both scenarios. We can see that including the Acc features did not introduce new, relevant information to the classifier, and therefore did not improve the recognition accuracy. Both scenarios were also evaluated using SVM with non-linear kernel (Radial Basis Function (RBF)), although linear SVM proved to be better, see Table 3. The parameters for the classifiers were determined by a grid search: $C = 1$ for the linear SVM, and $C = 10$ and $gamma = 0.024$ for the SVM-RBF.

Table 2: Recognition accuracy (%) for the FFD dataset using fusion of multiple feature sets with respect to different feature selection methods. Selections are made for each feature set separately.

	w/o sel.	PCA	AdaBoost	Lasso	Proposed
LTI + JD + JMHC	83.28	80.16	82.46	83.61	**86.31**
LTI + JD + JMHC + Acc	83.44	79.67	82.13	84.1	**85.66**

Table 3: Recognition accuracy (%) for the FFD dataset using the proposed feature selection method and fusion of multiple feature sets for linear SVM ($C = 1$) and SMV-RBF ($C = 10$ and $gamma = 0.024$).

	SVM LIN	SVM RBF
LTI + JD + JMHC	**86.31**	85.74
LTI + JD + JMHC + Acc	**85.66**	84.92

The obtained results are compared to state-of-the-art methods in Table 4, based on the comparison performed in [20]. The proposed method significantly improves previous results. Confusion matrix is presented in Table 5.

CNN-based approaches are not included in the comparison, because the FFD dataset does not contain RGB data. As explained in [20] RGB data are not well suited for sports action recognition, due to person extraction difficulties related to poor lighting and possible presence of multiple persons.

Table 4: Recognition accuracy for the FFD dataset - the proposed method compared to the state-of-the-art methods.

Method	Recognition accuracy (%)
EigenJoints [28]	29.89
MHI [56]	61.25
SkeletonNet [30]	64.36
C3D [57]	67.63
HON4D [33]	75.87
LOP/FTP [58]	76.14
LTI + JD [20]	81.49
LTI + JD + Acc [20]	83.59
LTI + JD + JMHC (ours)	**86.31**
LTI + JD + JMHC + Acc (ours)	**85.66**

Table 5: Confusion matrix for the proposed method (LTI + JD + JMHC) on the FFD dataset.

	R	IS	W/W	JS	SF	SB
R	**85.27**	12	1.82	0.91	-	-
IS	10.99	**71.64**	5.55	11.82	-	-
W/W	4.55	18.18	**77.27**	-	-	-
JS	-	13.64	-	**86.36**	-	-
SF	-	-	-	-	**100**	-
SB	-	-	-	-	2.68	**97.32**

5.3. UTD-MHAD results

Depth map and skeleton-based features for the UTD-MHAD dataset were computed for all available joints, separately in 3 orthogonal planes. Acc features include 3D data in a single feature vector. Experimental results demonstrated, that LTI features were not effective for this dataset, therefore only JD, JMHC and Acc features are considered.

Evaluation of the proposed feature selection algorithm was performed by comparison with reference methods (without selection, PCA, AdaBoost, Lasso).

The obtained results, which include the recognition accuracy and the number of selected features, are presented in Table 6. Linear SVM is used for the classification, with parameter $C = 1$ determined by a grid search. JD and JMHC features are evaluated separately for each plane. The proposed method obtains the best results in four out of seven cases, whereas for two of them AdaBoost provides similar accuracy. PCA is the most effective method in three cases, which indicates significant data redundancy, particularly in the case of JMHC features. This is due to the features being computed for all joints, even though in many actions only a relatively small subset of joints undergoes movement. For instance, in the *wave* action, only the motion of joints of one arm is relevant. It is worth noting that the proposed algorithm is only slightly less effective than PCA for two JMHC feature sets, while still being significantly more effective in comparison to other methods.

Table 6: Recognition accuracy (%) for the UTD-MHAD dataset using linear SVM ($C = 1$) and different feature selection methods on different feature sets.

		w/o sel.	PCA	AdaBoost	Lasso	Proposed
JD_{xy}	#feat	2640	100	500	900	600
	acc.	82.79	85.58	84.88	85.81	**86.05**
JD_{xz}	#feat	2640	200	1000	900	700
	acc.	84.19	75.81	**85.58**	85.35	**85.58**
JD_{yz}	#feat	2640	100	2000	1750	200
	acc.	83.95	79.07	**86.98**	85.12	**86.98**
JMHC_{xy}	#feat	16800	400	750	2000	2000
	acc.	63.02	**80**	66.51	72.56	78.14
JMHC_{xz}	#feat	16800	400	1250	600	800
	acc.	63.02	73.95	68.14	74.19	**82.09**
JMHC_{yz}	#feat	16800	400	1250	2500	4500
	acc.	64.19	**73.95**	65.12	68.84	71.63
Acc	#feat	168	100	120	140	150
	acc.	78.14	**79.53**	77.91	79.07	78.6

Evaluation of the feature selection in the proposed fusion scenario was performed next, see results in Table 7. Two scenarios were considered - using only Kinect-based features and with additional Acc features. In both scenarios the proposed method outperforms the algorithms used for comparison. We can also observe, that including Acc data provides better results. The recognition accuracy was further improved, by employing SVM with RBF kernel, see Table 8. The following values for the classifier parameters were determined by a grid search: $C = 1$ for the linear SVM, $C = 10$ and $gamma = 0.03$ for the SVM-RBF.

Table 7: Recognition accuracy (%) for the UTD-MHAD dataset using fusion of multiple feature sets with respect to different feature selection methods. Selections are made for each feature set separately.

	w/o sel.	PCA	AdaBoost	Lasso	Proposed
JD + JMHC	90.89	92.29	91.12	91.82	**92.76**
JD + JMHC + Acc	90.89	93.39	92.99	92.52	**94.39**

Table 8: Recognition accuracy (%) for the UTD-MHAD dataset using the proposed feature selection method and fusion of multiple feature sets for linear SVM ($C = 1$) and SVM-RBF ($C = 10$ and $gamma = 0.03$).

	SVM LIN	SVM RBF
JD + JMHC	92.76	**93.93**
JD + JMHC + Acc	94.39	**94.91**

Comparison of the proposed algorithm with state-of-the-art methods is presented in Table 9. Modalities used in each work are included in the above mentioned table. The proposed method outperforms all depth, skeleton and inertial-based methods. Also, it is only slighty worse (94.91% compared to 95.11%) than the best VGG-16 based algorithm, which employs RGB, depth and skeleton modalities, as well as uses a deep CNN [59]. It is worth noting, that our method operates in real-time. The average processing time of a single frame is 17 ms (measured on a machine with Intel Core i5 2.5 GHz CPU).

When using the extracted features, the average time needed for classification of an action is 2 ms. On the other hand, large CNN architectures are unable to operate in real-time and for this reason smaller architectures are proposed [60]. VGG-F [59], which is a faster, smaller version of the VGG-16 network, obtains lower recognition accuracy than our method, even though it employs RGB data (see Table 9).

Confusion matrix for the obtained results is presented in Table 9. Out of 27 actions, 21 are recognized without any error. Pairs of similar actions, such as *draw circle CW* and *draw circle CCW* were the most difficult to distinguish.

Table 9: Recognition accuracy for the UTD-MHAD dataset - the proposed method compared to the state-of-the-art methods.

Method	Acc. (%)	Modalities
DMM-CRC [23]	79.10	Depth + Inertial
GF + LF [61]	84.89	Depth + Skeleton
SD-SR [62]	86.12	Skeleton
JTM + CNN [63]	87.90	Skeleton
DMM-CT-HOG-LBP-EOH [64]	88.40	Depth
DMM-CRC-LOGP [65]	91.50	Depth + Skeleton + Inertial
TPM-LLC-BoA [66]	93.02	Skeleton
MDACC [67]	93.26	Depth + Skeleton + Inertial
VGG-F [59]	94.60	RGB + Depth + Skeleton
VGG-16 [59]	**95.11**	**RGB + Depth + Skeleton**
JD + JMHC (ours)	**93.93**	**Depth + Skeleton**
JD + JMHC + Acc (ours)	**94.91**	**Depth + Skeleton + Inertial**

Table 10: Confusion matrix for the proposed method (JD + JMHC + Acc) on the UTD-MHAD dataset.

	a1	a2	a3	a4	a5	a6	a7	a8	a9	a10	a11	a12	a13	a14	a15	a16	a17	a18	a19	a20	a21	a22	a23	a24	a25	a26	a27
swipe left	**100**	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
swipe right	-	**100**	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
wave	-	-	**100**	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
clap	-	-	-	**100**	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
throw	18.75	-	-	-	**81.25**	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
arm cross	-	-	-	-	-	**100**	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
basketball shoot	-	-	-	-	-	-	**100**	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
draw x	-	-	-	-	-	-	-	**100**	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
draw circle CW	-	-	-	-	-	-	-	-	**100**	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
draw circle CCW	-	-	-	-	-	-	-	-	43.75	**56.25**	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
draw triangle	-	-	6.25	-	-	-	-	-	-	12.5	**81.25**	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
bowling	-	-	-	-	-	-	-	-	-	-	-	**100**	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
boxing	-	-	-	-	-	-	-	-	-	-	-	-	**100**	-	-	-	-	-	-	-	-	-	-	-	-	-	-
baseball swing	-	-	-	-	-	-	-	-	-	-	-	-	-	**100**	-	-	-	-	-	-	-	-	-	-	-	-	-
tennis swing	-	-	-	-	-	-	-	-	-	-	-	-	-	-	**100**	-	-	-	-	-	-	-	-	-	-	-	-
arm curl	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	**100**	-	-	-	-	-	-	-	-	-	-	-
tennis serve	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	**100**	-	-	-	-	-	-	-	-	-	-
push	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	**100**	-	-	-	-	-	-	-	-	-
knock	-	-	6.25	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	**93.75**	-	-	-	-	-	-	-	-
catch	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	25	**75**	-	-	-	-	-	-	-
pickup&throw	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	**100**	-	-	-	-	-	-
jog	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	**75**	25	-	-	-	-
walk	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	**100**	-	-	-	-
sit to stand	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	**100**	-	-	-
stand to sit	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	**100**	-	-
lunge	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	**100**	-
squat	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	**100**

6. Conclusions

In this paper we proposed a novel depth map and skeleton-based action descriptor, which is able to improve action representation in a multimodal scenario and therefore provides better action recognition accuracy. The proposed descriptor considers motion patterns relative to each joint, by employing log-polar histograms to quantify the observed motion. The temporal structure of the action is captured by using motion differences that are computed between multiple frames, as well as by employing time segments for creating the final feature vector. An effective feature selection method is also proposed, which ranks features based on their inter-class discriminative strength, but also reduces redundancy in the final selected subset.

The proposed algorithms are evaluated on two publicly available datasets. Our feature selection method outperforms state-of-the-art algorithms - PCA, AdaBoost and Lasso. The results obtained in action recognition on both datasets are better than for relevant methods from the literature, when using the same modalities. For the UTD-MHAD dataset, recognition accuracy is only slightly lower than the best approach, which employs additional channel with RGB data and a large CNN. For both datasets, using multiple feature sets computed from the depth and skeleton data provided better results compared to using separate feature sets, which indicates that the employed decision-level fusion is effective. Using even more feature sets for the fusion is an interesting subject, worth investigating in future work.

Acknowledgements

This work was supported by Polish National Science Center (NCN) under a research grant 2014/15/B/ST6/02808 and by the funds of Polish Ministry of Science and Higher Education assigned to AGH University of Science and Technology.

References

[1] F. Zhu, L. Shao, J. Xie, and Y. Fang, "From handcrafted to learned representations for human action recognition: A survey," *Image and Vision Computing*, vol. 55, pp. 42–52, 2016.

[2] P. Skalski, R. Tamborini, A. Shelton, M. Buncher, and P. Lindmark, "Mapping the road to fun: Natural video game controllers, presence, and game enjoyment," *New Media & Society*, vol. 13, no. 2, pp. 224–242, 2011.

[3] V. Fernandez-Cervantes, N. Neubauer, B. Hunter, E. Stroulia, and L. Liu, "VirtualGym: A Kinect-based system for seniors exercising at home," *Entertainment Computing*, vol. 27, pp. 60–72, 2018.

[4] D. González-Ortega, F. Díaz-Pernas, M. Martínez-Zarzuela, and M. Antón-Rodríguez, "A Kinect-based system for cognitive rehabilitation exercises monitoring," *Computer Methods and Programs in Biomedicine*, vol. 113, no. 2, pp. 620–631, 2014.

[5] A. B. Mabrouk and E. Zagrouba, "Abnormal behavior recognition for intelligent video surveillance systems: A review," *Expert Systems with Applications*, vol. 113, pp. 480–491, 2018.

[6] W. Lin, Y. Zhou, H. Xu, J. Yan, M. Xu, J. Wu, and Z. Liu, "A tube-and-droplet-based approach for representing and analyzing motion trajectories," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 8, pp. 1489–1503, 2017.

[7] M. Rawashdeh, M. G. Al Zamil, S. Samarah, M. S. Hossain, and G. Muhammad, "A knowledge-driven approach for activity recognition in smart homes based on activity profiling," *Future Generation Computer Systems*, 2017.

[8] M. Hamaya, T. Matsubara, T. Noda, T. Teramae, and J. Morimoto, "Learning assistive strategies for exoskeleton robots from user-robot physical interaction," *Pattern Recognition Letters*, vol. 99, pp. 67–76, 2017.

[9] G. Thomas, R. Gade, T. B. Moeslund, P. Carr, and A. Hilton, "Computer vision for sports: current applications and research topics," *Computer Vision and Image Understanding*, vol. 159, pp. 3–18, 2017.

[10] S. Kasiri, C. Fookes, S. Sridharan, and S. Morgan, "Fine-grained action recognition of boxing punches from depth imagery," *Computer Vision and Image Understanding*, vol. 159, pp. 143–153, 2017.

[11] B. Reily, H. Zhang, and W. Hoff, "Real-time gymnast detection and performance analysis with a portable 3D camera," *Computer Vision and Image Understanding*, vol. 159, pp. 154–163, 2017.

[12] F. Malawski and B. Kwolek, "Real-time action detection and analysis in fencing footwork," in *2017 40th International Conference on Telecommunications and Signal Processing (TSP)*. IEEE, 2017, pp. 520–523.

[13] D. Weinland, R. Ronfard, and E. Boyer, "A survey of vision-based methods for action representation, segmentation and recognition," *Computer Vision and Image Understanding*, vol. 115, no. 2, pp. 224–241, 2011.

[14] L. L. Presti and M. La Cascia, "3D skeleton-based human action classification: A survey," *Pattern Recognition*, vol. 53, pp. 130–147, 2016.

[15] H. Sarbolandi, D. Lefloch, and A. Kolb, "Kinect range sensing: Structured-light versus Time-of-Flight Kinect," *Computer Vision and Image Understanding*, vol. 139, pp. 1–20, 2015.

[16] K. Altun, B. Barshan, and O. Tunçel, "Comparative study on classifying human activities with miniature inertial and magnetic sensors," *Pattern Recognition*, vol. 43, no. 10, pp. 3605–3620, 2010.

[17] T. T. Ngo, Y. Makihara, H. Nagahara, Y. Mukaigawa, and Y. Yagi, "Similar gait action recognition using an inertial sensor," *Pattern Recognition*, vol. 48, no. 4, pp. 1289–1301, 2015.

[18] Z. Gao, S. Li, Y. Zhu, C. Wang, and H. Zhang, "Collaborative sparse representation leaning model for RGBD action recognition," *Journal of Visual Communication and Image Representation*, vol. 48, pp. 442–452, 2017.

[19] C. Chen, R. Jafari, and N. Kehtarnavaz, "Improving human action recognition using fusion of depth camera and inertial sensors," *IEEE Transactions on Human-Machine Systems*, vol. 45, no. 1, pp. 51–61, 2015.

[20] F. Malawski and B. Kwolek, "Recognition of action dynamics in fencing using multimodal cues," *Image and Vision Computing*, vol. 75, pp. 1–10, 2018.

[21] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 3551–3558.

[22] W. Lin, C. Zhang, K. Lu, B. Sheng, J. Wu, B. Ni, X. Liu, and H. Xiong, "Action recognition with coarse-to-fine deep feature integration and asynchronous fusion," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[23] C. Chen, R. Jafari, and N. Kehtarnavaz, "UTD-MHAD: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor," in *IEEE International Conference on Image Processing (ICIP)*. IEEE, 2015, pp. 168–172.

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

[25] M. Kepski and B. Kwolek, "Detecting human falls with 3-axis accelerometer and depth sensor," in *Engineering in Medicine and Biology Society (EMBC), 36th Annual International Conference of the IEEE*. IEEE, 2014, pp. 770–773.

[26] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, and R. Bajcsy, “Sequence of the most informative joints (SMIJ): A new representation for human skeletal action recognition,” *Journal of Visual Communication and Image Representation*, vol. 25, no. 1, pp. 24–38, 2014.

[27] C. Ellis, S. Z. Masood, M. F. Tappen, J. J. LaViola, and R. Sukthankar, “Exploring the trade-off between accuracy and observational latency in action recognition,” *International Journal of Computer Vision*, vol. 101, no. 3, pp. 420–436, 2013.

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

[29] Z. Wang, C. Zhang, W. Luo, and W. Lin, “Key joints selection and spatiotemporal mining for skeleton-based action recognition,” in *25th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2018, pp. 3458–3462.

[30] Q. Ke, S. An, M. Bennamoun, F. Sohel, and F. Boussaid, “SkeletonNet: Mining deep part features for 3-D action recognition,” *IEEE Signal Processing Letters*, vol. 24, no. 6, pp. 731–735, 2017.

[31] “learning clip representations for skeleton-based 3d action recognition.”

[32] J. Liu, A. Shahroudy, G. Wang, L.-Y. Duan, and A. C. Kot, “SSNet: Scale selection network for online 3D action prediction,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 8349–8358.

[33] O. Oreifej and Z. Liu, “HON4D: Histogram of oriented 4D normals for activity recognition from depth sequences,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013, pp. 716–723.

[34] X. Yang, C. Zhang, and Y. Tian, “Recognizing actions using depth motion maps-based histograms of oriented gradients,” in *Proceedings of the 20th*

ACM International Conference on Multimedia. ACM, 2012, pp. 1057–1060.

[35] A. W. Vieira, E. R. Nascimento, G. L. Oliveira, Z. Liu, and M. F. Campos, "STOP: Space-time occupancy patterns for 3D action recognition from depth map sequences," in *Iberoamerican Congress on Pattern Recognition.* Springer, 2012, pp. 252–259.

[36] Z. Liu, C. Zhang, and Y. Tian, "3D-based Deep Convolutional Neural Network for action recognition with depth sequences," *Image and Vision Computing*, vol. 55, pp. 93–100, 2016.

[37] A. M. Sabatini, C. Martelloni, S. Scapellato, and F. Cavallo, "Assessment of walking features from foot inertial sensing," *IEEE Transactions on Biomedical Engineering*, vol. 52, no. 3, pp. 486–494, 2005.

[38] O. D. Lara, M. A. Labrador *et al.*, "A survey on human activity recognition using wearable sensors." *IEEE Communications Surveys and Tutorials*, vol. 15, no. 3, pp. 1192–1209, 2013.

[39] J. Parkka, M. Ermes, P. Korpipaa, J. Mantyjarvi, J. Peltola, and I. Korhonen, "Activity classification using realistic data from wearable sensors," *IEEE Transactions on Information Technology in Biomedicine*, vol. 10, no. 1, pp. 119–128, 2006.

[40] L. Bao and S. S. Intille, "Activity recognition from user-annotated acceleration data," in *International Conference on Pervasive Computing.* Springer, 2004, pp. 1–17.

[41] Y. Hanai, J. Nishimura, and T. Kuroda, "Haar-like filtering for human activity recognition using 3D accelerometer," in *IEEE 13th Digital Signal Processing Workshop and 5th IEEE Signal Processing Education Workshop (DSP/SPE).* IEEE, 2009, pp. 675–678.

[42] 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, pp. 37–49, 2016.

[43] F. Malawski and B. Kwolek, “Classification of basic footwork in fencing using accelerometer,” in *Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA)*, Sept 2016, pp. 51–55.

[44] C. Chen, R. Jafari, and N. Kehtarnavaz, “A survey of depth and inertial sensor fusion for human action recognition,” *Multimedia Tools and Applications*, vol. 76, no. 3, pp. 4405–4425, 2017.

[45] B. Delachaux, J. Rebetez, A. Perez-Uribe, and H. F. S. Mejia, “Indoor activity recognition by combining one-vs.-all neural network classifiers exploiting wearable and depth sensors,” in *International Work-Conference on Artificial Neural Networks.* Springer, 2013, pp. 216–223.

[46] B. Kwolek and M. Kepski, “Improving fall detection by the use of depth sensor and accelerometer,” *Neurocomputing*, vol. 168, pp. 637–645, 2015.

[47] S. Belongie, J. Malik, and J. Puzicha, “Shape matching and object recognition using shape contexts,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 4, pp. 509–522, Apr 2002.

[48] G. Lu and M. Kudo, “Learning action patterns in difference images for efficient action recognition,” *Neurocomputing*, vol. 123, pp. 328–336, 2014.

[49] I. Guyon and A. Elisseeff, “An introduction to variable and feature selection,” *Journal of Machine Learning Research*, vol. 3, no. Mar, pp. 1157–1182, 2003.

[50] P. Viola and M. J. Jones, “Robust real-time face detection,” *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004.

[51] F. Malawski, B. Kwolek, and S. Sako, “Using Kinect for facial expression recognition under varying poses and illumination,” in *International Conference on Active Media Technology.* Springer, 2014, pp. 395–406.

[52] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

[53] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 2, no. 4, pp. 433–459, 2010.

[54] A. Jain and D. Zongker, "Feature selection: Evaluation, application, and small sample performance," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 2, pp. 153–158, 1997.

[55] "Weka Data Mining Software," https://www.cs.waikato.ac.nz/ml/weka/, last access on Jan 2019.

[56] A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 3, pp. 257–267, 2001.

[57] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 4489–4497.

[58] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Mining actionlet ensemble for action recognition with depth cameras," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2012, pp. 1290–1297.

[59] P. Khaire, P. Kumar, and J. Imran, "Combining CNN streams of RGB-D and skeletal data for human activity recognition," *Pattern Recognition Letters*, vol. 115, pp. 107–116, 2018.

[60] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," *arXiv preprint arXiv:1405.3531*, 2014.

[61] E. Escobedo and G. Camara, “A new approach for dynamic gesture recognition using skeleton trajectory representation and histograms of cumulative magnitudes,” in *29th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI).* IEEE, 2016, pp. 209–216.

[62] Y. Annadani, D. Rakshith, and S. Biswas, “Sliding dictionary based sparse representation for action recognition,” *arXiv preprint arXiv:1611.00218*, 2016.

[63] P. Wang, W. Li, C. Li, and Y. Hou, “Action recognition based on joint trajectory maps with convolutional neural networks,” *Knowledge-Based Systems*, vol. 158, pp. 43–53, 2018.

[64] M. F. Bulbul, Y. Jiang, and J. Ma, “DMMs-based multiple features fusion for human action recognition,” *International Journal of Multimedia Data Engineering and Management (IJMDEM)*, vol. 6, no. 4, pp. 23–39, 2015.

[65] C. Chen, R. Jafari, and N. Kehtarnavaz, “A real-time human action recognition system using depth and inertial sensor fusion,” *IEEE Sensors Journal*, vol. 16, no. 3, pp. 773–781, 2016.

[66] N. E. D. Elmadany, Y. He, and L. Guan, “Human gesture recognition via bag of angles for 3D virtual city planning in CAVE environment,” in *IEEE 18th International Workshop on Multimedia Signal Processing (MMSP).* IEEE, 2016, pp. 1–5.

[67] N. E. D. El Madany, Y. He, and L. Guan, “Human action recognition via multiview discriminative analysis of canonical correlations,” in *Image Processing (ICIP), 2016 IEEE International Conference on.* IEEE, 2016, pp. 4170–4174.