Ann Oper Res (2018) 265:223–239
https://doi.org/10.1007/s10479-016-2308-z

Human actions recognition from motion capture recordings using signal resampling and pattern recognition methods

Tomasz Hachaj[1] · **Marek R. Ogiela**[2] · **Katarzyna Koptyra**[2]

Published online: 15 September 2016

Abstract In this paper we will experimentally prove that after recalculating the motion capture (MoCap) data to position-invariant representation it can be directly used by classifier to successfully recognize various actions types. The assumption on classifier is that it is capable to deal with objects that are described by hundreds of numeric values. The second novelty of this paper is application of neural network trained with the parallel stochastic gradient descent, Random Forests and Support Vector Machine with Gaussian radial basis kernel to perform classification task on gym exercises and karate techniques MoCap datasets. We have tested our approach on two datasets using k-fold cross-validation method. Depending of the dataset we have obtained averaged recognition rate from 100 to 97 %. Our results presented in this work give very important hints for developing similar actions recognition systems because proposed features selection and classification setup seems to guarantee high efficiency and effectiveness.

Keywords Actions recognition · Neural network · Support vector machine · Random forest · Motion capture · Kinect

✉ Tomasz Hachaj
tomekhachaj@o2.pl

Marek R. Ogiela
mogiela@agh.edu.pl

Katarzyna Koptyra
kkoptyra@agh.edu.pl

[1] Institute of Computer Science, Pedagogical University of Krakow, 2 Podchorazych Ave, 30-084 Krakow, Poland

[2] Cryptography and Cognitive Informatics Research Group, AGH University of Science and Technology, 30 Mickiewicza Ave, 30-059 Krakow, Poland

1 Introduction

Human actions recognition is challenging and up-to-date problem that appears in many practical applications like computer games, security monitoring or smart home technologies. In this section we will present state-of-the-art review in actions recognition methods and our motivation for writing this paper.

1.1 State-of-the-art in actions recognition

Nearly each actions recognition framework proposed in the literature introduces its own feature selection method. Neural networks (NN) are among pattern recognition methods that were commonly reported to be used for actions recognition and human pose estimation Jiu et al. (2012), Li et al. (2015), Chen et al. (2015), Charalampous and Gasteratos (2014). Also paper Li et al. (2014) proposes a framework that combines Fast HOG3D description and self-organization feature map (SOM) network for actions recognition from unconstrained videos, bypassing the demanding preprocessing such as human detection, tracking or contour extraction. Support vector machines (SVM) are also among supervised classification method used for actions recognition Liu et al. (2013a), Díaz-Más et al. (2012), Mahbub et al. (2014), Shen et al. (2015), Cao et al. (2014), Chen et al. (2015), Ji et al. (2014), Bilen et al. (2014), Omidyeganeh et al. (2013), Nasiri et al. (2014), Zhen et al. (2014), Wu et al. (2014). The different class of pattern classification methods designed for actions recognition is that which uses rule-based descriptions and reasoning modules. Among those is Gesture Description Language Hachaj and Ogiela (2014) that uses unsupervised R-GDL training Hachaj and Ogiela (2014, 2015a, b) for automatic rules generation. GDL can also be use as online video segmentation method that prepares the input signal to other classification methods like hidden Markov model (HMM) Hachaj et al. (2015a, b). Paper Rincón et al. (2013) proposed methodology is decomposed into two stages. First, a bag-of-words gives a first estimate of action classification from video sequences, by performing an image feature analysis. Those results are afterward passed to a common-sense reasoning system, which analyses, selects and corrects the initial estimation yielded by the machine learning algorithm. This second stage resorts to the knowledge implicit in the rationality that motivates human behavior. Some action classification tasks can be solved with simple naive Bayes nearest-neighbor method Liu et al. (2013b) and Yang and Tian (2014). Random forests (RF) approach is popular method utilized in process of segmentation and recognition of actions Zhu et al. (2013), Jiang et al. (2013), Saito and Nishiyama (2015), Liu et al. (2014), Burghouts et al. (2014), Burghouts et al. (2013), Chen and Guo (2015), Jiang et al. (2013). SVM and RF are very flexible approaches that have many important applications and can operate on objects described by various features sets Fan and Chaovalitwongse (2010), Yahav and Shmueli (2014). Among features and features selection methods that are often applied for human actions recognition there are methods like optical flow Liu et al. (2013b), Mahbub et al. (2014), Jiang et al. (2013), Liu et al. (2014) various dimensionality reduction techniques like PCA, 2D-PCA, LDA, Díaz-Más et al. (2012), bag-of-words framework Shen et al. (2015), Cao et al. (2014), Nasiri et al. (2014), Burghouts et al. (2013), probability distributions - based features Chen et al. (2015), Ji et al. (2014) or 3D wavelet transform Omidyeganeh et al. (2013). There are also a number of pattern recognition methods that are less commonly used in human actions recognition tasks. We can mention regularized multi-task learning Guo and Chen (2015), papers Hachaj and Ogiela (2014, 2015a, b) models actions with multivariate continuous hidden Markov model classifier, dynamic time warping, canonical time warping Vrigkas et al. (2014). In paper Jiang et al. (2015) and Liu et al. (2015) feature sets are evaluated using a Conditional

Random Fields linear (CRFs). In paper Pazhoumand-Dar et al. (2015) author uses longest common subsequence (LCSS) algorithm to assign action represented by body joints derived features to proper class.

The state of the art review on recent developments in deep learning and unsupervised feature learning for time-series problems can be found in Längkvist et al. (2014) while Ziaeefard and Bergevin (2015) presents an overview of state-of-the-art methods in activity recognition using semantic features.

1.2 Our motivation for writing this paper

As can be seen in above state-of-the-art review one of the most challenging stage of actions recognition is appropriate features selection that enables to extract the movements characteristics from video sequence. However up-to-date multimedia depth cameras like for example Kinect controllers enables relatively cheap registration of video stream that can be then used for extraction of human posture and so called skeleton. This approach is marker-less MoCap. There are number of methods that are capable for this type of extraction and body joints tracking Papadopoulos et al. (2014), Shotton et al. (2013), Coleca et al. (2013). The tracked features consisted of so called body joints are valuable source of information that does not require much further processing to be used by classifier. State-of-the-art papers however even when dealing with skeleton data processes it with additional methods making the output data dependent to many additional parameters. Those parameters values are often dependent on processing model and might differ between actions to which we want to apply them. In fact the feature set that describes an action has one crucial demand—it has to be invariant to relative position of observed user to camera. In this paper we will experimentally prove that after recalculating the MoCap data to position-invariant representation it can be directly used by classifier to successfully recognize various actions types. The assumption on classifier is that it is capable to deal with objects that are described by hundreds of numeric values. For example up-to-date implementation of parallel stochastic gradient descent training method Recht et al. (2011) allows to relatively quickly train NN that is dependent on hundreds of thousands synaptic weights.

The second novelty of this paper is application of NN trained with the parallel stochastic gradient descent, Random Forests and SVM with Gaussian radial basis kernel to perform classification task on gym exercises and karate techniques MoCap datasets. The original MoCap data consisted of 20 or 25 time-varying three-dimensional body joints coordinates acquired with Kinect (appropriately Kinect 2) controller is preprocessed to 9-dimensional angle-based time-varying features set, 15-dimensional or 16-dimensional distance based feature set. The data is resampled to the uniform length with cubic spline interpolation after which each action is represented by 60 samples and eventually 540 (60×9), 900 (60×15) or 960 (60×16)-dimensional variables are presented to the classifier. We have tested our approach on two datasets using k-fold cross-validation method. First dataset introduced in Hachaj and Ogiela (2015a) consists of recordings of 14 participants that perform nine types of popular gym exercises (totally 770 actions samples). The second dataset is extended version of one introduced in Hachaj et al. (2015a). It consists of recordings of 6 participants that perform sixteen types of karate techniques (totally 1996 actions samples). In the following sections we will present the dataset we have used in our experiment, feature selection methodology and classification methods. Later we will also discuss the obtained results and present goals for future researches.

2 Material and methods

In this section we will present the dataset, features selection procedure and classifiers we have used in our experiment.

2.1 Dataset and features selection

The launching of Microsoft Kinect with skeleton tracking technique opens up new potentials for skeleton based human actions recognition. However, the 3D human skeletons, generated via skeleton tracking from the depth map sequences, are generally very noisy and unreliable what makes actions recognition a challenging task Jiang et al. (2015). Despite the fact that Kinect was initially designed to be a game controller, its potential as cheap general purpose depth camera was quickly noticed Hachaj et al. (2015b).

To gather the dataset for evaluation of proposed methodology we have utilized Microsoft Kinect v1 for the gym exercises dataset and Microsoft Kinect v2 for karate techniques dataset. Those datasets were prepared using different hardware because in time when gym dataset was recorded Kinect v2 was not yet available. According to research Hachaj et al. (2015b) Kinect v2 controller and Kinect v2 SDK is capable to generate more reliable data for classification in competition to Kinect v1 so second dataset was recorded using the newer hardware. The Kinect SDK software library for Kinect v1 is capable to segment and track 20 joints on human body with acquisition frequency of 30Hz while SDK for Kinect v2 segments and tracks 25 joints with the same frequency. The tracking is marker-less procedure. We have used those joints to produce camera position invariant representation of action because the dependence on the camera position virtually prevents method from being usable in real-world scenario. In our angle-based representation (Fig. 1a) the vertices of angles are positioned either in some important for movements analysis body joints (like elbows—angle 1 and 2, shoulders—angle 3 and 4, knees—angle 6 and 7) or angles measure position of limbs relatively to each other or relatively to torso. The second type of angles we utilized are angle defined between forearms (angle 5), angle between vector defined by joint between shoulders–joint between hips and thighs (angle 8 and 9). The same representation was used for both Kinect v1 and Kinect v2 datasets. The selection of this subset of all possible angles was among subset considered in Hachaj and Ogiela (2015b) for which HMM used their obtained high recognition rate. The second and third feature set was defined as Euclidean distances between central joint (in Fig. 1b, c) it is "spine" joint with index 0) and 15 other joints in (B) and 16 in (C). The joints we used are nearly all joints form Kinect SDK beside feet and hands joints that we skipped due to high inaccuracies of tracking of those body parts. The above joints representations were calculated to all frames of acquired actions recordings. In the next step the data is resampled to the uniform length with cubic spline interpolation after which each action is represented by the vector of the same size. The uniform length we choose was 60 frames per recording which was the smallest number of frames that was present among all actions recordings in both considered datasets. After this operation gym exercises dataset was represent by 540 variables (60×9—see Fig. 1a) or by 900 variables (60×15—see Fig. 1b). The karate techniques dataset was represented also by 540 variables (60×9) or by 960 variables (60×16—see Fig. 1c). All those features sets were evaluated separately in our experimental setup.

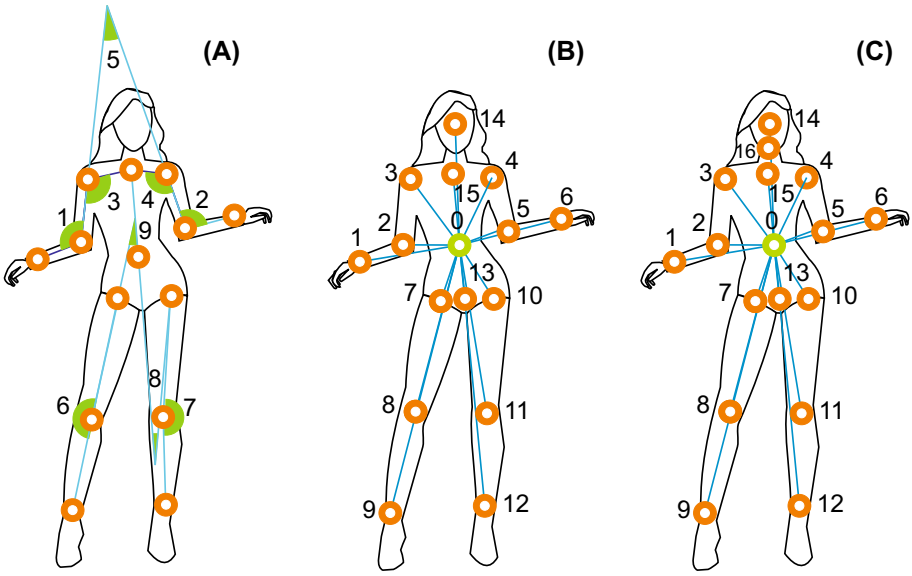

Fig. 1 This figure presents the three features sets we utilized in our experiment. The first one (**a**) is angle-based the next two (**b**) and (**c**) are distance-based

2.2 Neural network implementation

In our experiment we used multi-layer, feedforward neural networks Candel and Parmer (2015). It consists of many layers of interconnected neuron units: beginning with an input layer to match the feature space followed by a layer of nonlinearity and terminating with a classification layer to match the output space. For each training example j the objective is to minimize a loss function $L(W, B|j)$, where W is the collection $\{w_i\}_{1:N-1}$, W_i denotes the weight matrix connecting layers i and $i+1$ for a network of N layers; similarly B is the collection $\{b\}_{1:N-1}$, where b_i denotes the column vector of biases for layer $i+1$.

The training of NN for classification task is based on minimization of cross-entropy loss function Candel and Parmer (2015):

$$(W, B|j) = -\sum_{y\in O}\left(\ln(o_y^{(j)})\cdot t_y^{(j)} + ln(1-o_y^{(j)})\cdot(1-t_y^{(j)})\right) \tag{1}$$

where $o_y^{(j)}$ and $t_y^{(j)}$ are the predicted (target) output and actual output, respectively, for training example j, and $y$denote the output units and $O$the output layer.

For minimization of (1) stochastic gradient descent (SGD) method can be used which is an iteration procedure for each tanning example j LeCun et al. (2002):

$$\begin{cases} w_{jk} := w_{jk} - \alpha\frac{\partial L(W,B|j)}{\partial w_{jk}} \\ b := b_{jk} - \alpha\frac{\partial L(W,B|j)}{\partial b_{jk}} \end{cases} \tag{2}$$

where $w_{jk} \in W$ (weights), $b_{jk} \in B$ (biases).

To speed-up the training procedure, we used Hogwild, the lock-free parallelization scheme for SGD that has been published lately Recht et al. (2011).

The activation function in hidden layer might be a rectified linear function:

$$f(\alpha) = max(0, \alpha) \tag{3}$$

where:

$$\alpha = \sum_i w_i x_i + b \tag{4}$$

x_i and w_i denote the firing neuron's input values and their weights, respectively; α denotes the weighted combination.

In our experiment we have utilized fully connected NN. Input layer had 540, 900 or 960 neurons, depending on number of variables in features set. We have experimented with different number of neurons in hidden layer from 4 to 256. Activation function of neurons in hidden layer was (3). The input data for network is standardize to $N(0, 1)$.

2.3 Support vector machine implementation

Kernel-based learning methods use kernel function for mapping of the input data into a high dimensional feature space Karatzoglou et al. (2004). The further learning takes place in the feature space and the data points only appear inside dot products with other points. ("kernel trick") Schölkopf and Smola (2002). If a projection $\Phi : X \rightarrow H$ is used, the dot product $\Phi(\mathrm{x}) \circ \Phi(\mathrm{y})$ can be represented by a kernel function k:

$$\mathrm{k}(\mathrm{x}, \mathrm{y}) = \Phi(\mathrm{x}) \circ \Phi(\mathrm{y}) \tag{5}$$

which is computationally simpler than explicitly projecting x and y into the feature space H Karatzoglou et al. (2004). Support vector machines Vapnik (1998) have gained prominence in the field of machine learning and pattern classification and regression. The solutions to classification and regression problems such as the SVM are linear functions in the feature space:

$$f(x) = w^T \Phi(x) \tag{6}$$

where $w \in F$ is a weight vector. If the weight vector w can be expressed as a linear combination of the training points the kernel trick can be exploited:

$$f(x) = \sum_{i=1}^{n} \alpha_i k(x_i, x) \tag{7}$$

In the case of the 2-norm Soft Margin classification the optimization problem during classifier learning takes the form:
Minimize:

$$t(w, \zeta) = \frac{1}{2} \cdot \|w\|^2 + \frac{C}{m} \cdot \sum_{i=1}^{m} \zeta_i \tag{8}$$

Subject to:

$$y((x_i \circ w) + b) \geq 1 - \zeta_i, \zeta_i \geq 0, (1 = 1, \ldots, m) \tag{9}$$

The classification problems that include more than two classes (multi-class) a one-against-one Knerr et al. (1990) or pairwise classification method Kreßel (1999) is used. In our research we use Gaussian radial basis kernel:

$$k(x, x') = e^{(-\sigma \cdot \|x - x'\|^2)}, \sigma = 0.1, \tag{10}$$

Table 1 This table presents the quantities of gym exercises in dataset we used in our experiment

Dataset	bwll	bwlr	bws	dbc	jj	sll	slr	sdur	tdk
W1	10	10	10	10	10	10	10	10	10
W2	5	5	5	5	11	5	5	5	6
M1	13	10	12	11	9	10	7	12	10
M2	10	10	10	10	12	12	9	10	10
M3	10	10		10	10	8	10	10	10
W3			10						
M4	7	5		5	5		5	5	5
W4			5						
M5	10	10		10	10	5	10	10	10
M6			10						
M7	10	10	10	10	10	10	10	10	10
M8	5	5		5	5	6	5	5	5
M9			5						
M10	10	10	10	10	10	10	10	10	10
Sum	90	85	87	86	92	76	81	87	86

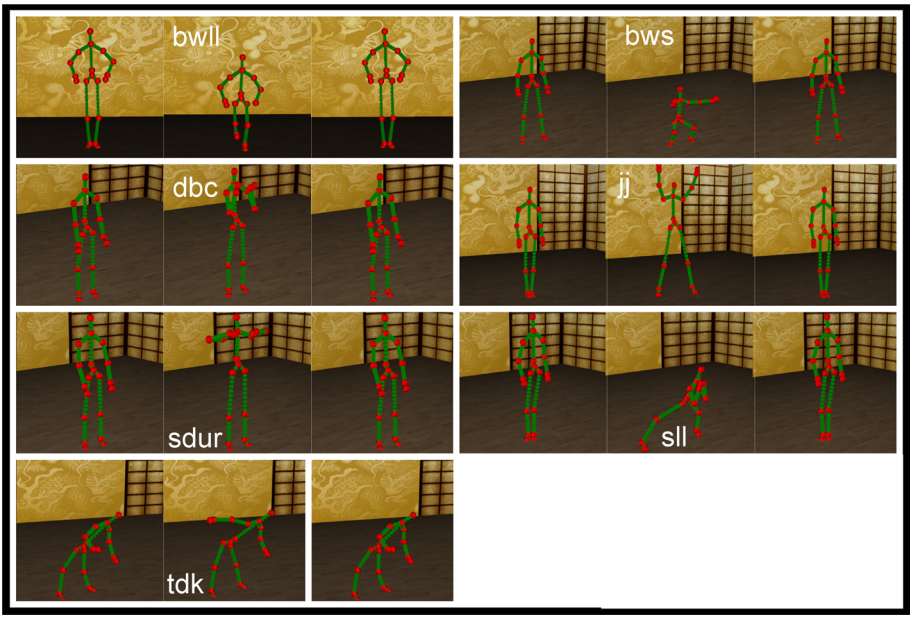

Fig. 2 This figure presents important phases of actions from the gym exercises dataset. The skeleton data is visualized in 3D virtual environment

2.4 Random forests implementation

Random forests are a combination of tree predictors. For all trees in the forest each tree depends on the values of a random vector sampled independently and with the same distribution. As the number of trees in becomes large the generalization error for forests decreases.

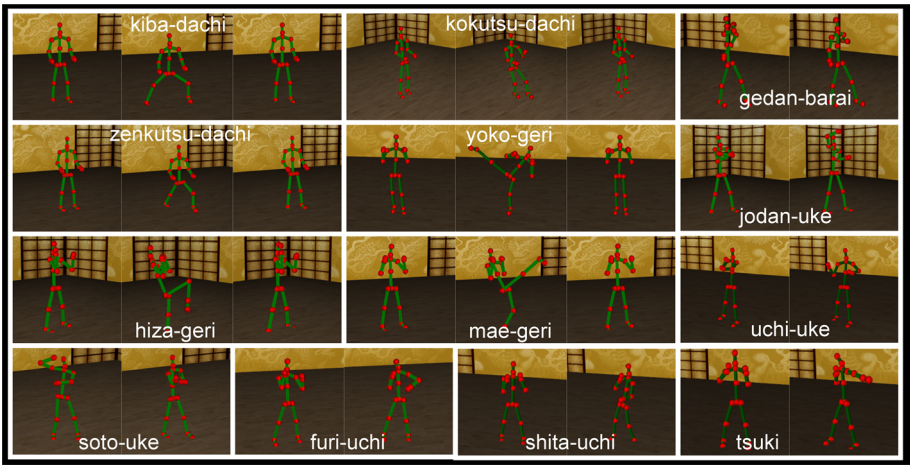

Fig. 3 This figure presents important phases of actions from the karate techniques dataset. The skeleton data is visualized in 3D virtual environment

That error depends on the strength of the individual trees in the forest and the correlation between them Breiman (2001). Each tree uses only random sample of training data and captures only a part of overall information. This is called a bagging procedure. The second randomized procedure is features selection during determining the best split. In H2O (2015) implementation we used in our experiment tree selects randomly subset of features of size square root of all features. The simplest random forest with random features is formed by selecting at random, at each node, a small group of input variables to split on. The tree growth uses CART methodology Breiman et al. (1984).

3 Results

The gym exercises dataset was used in earlier work Hachaj and Ogiela (2015a). It consists of recordings of 14 participants, 10 men (M1–M10) and 4 women (W1–W4), numbers defines id of a participant (see Table 1). The users were ask to perform: body weight lunge left (bwll), body weight lunge right (bwlr), body weight squat (bws), dumbbell bicep curl (dbc), jumping jacks (jj), side lunges left (sll), side lunges right (slr), standing dumbbell upright row (sdur), tricep dumbbell kickback (tdk). In Table 1 we have presented quantities of actions of a given type that was performed by each person. Total number of samples was 770. The visualization of important phases of actions from the gym exercises dataset is presented in Fig. 2.

The karate techniques dataset is extension of dataset we used in earlier work Hachaj et al. (2015a). The dataset consisted of MoCap recordings of six volunteers including multiple champion of Kumite Knockdown Oyama karate. We recorded four types of defense techniques (gedan-barai, jodan-uke, soto-uke and uchi-uke) three types of kicks (hiza-geri, mae-geri and yoko-geri) and three stands (kiba-dachi, kokutsu-dachi and zenkutsu-dachi). The stands were preceded by fudo-dachi and were also evaluated as actions (not as static body positions). Kicks were done with right foot and blocks were done with right hand. The original dataset was extended by three types of punches: furi-uchi, shita-uchi and tsuki. Punches were done with right and left hand separately. In Fig. 3 we present important stages of karate techniques we have evaluated. Total number of samples was 1996 (see Table 2).

Table 2 This table presents quantities of MoCap movement samples we have gathered from six volunteers

	gedan-barai	hiza-geri	jodan-uke	kiba-dachi	kokutsu-dachi	mae-geri	soto-uke	uchi-uke	yoko-geri	zenkutsu-dachi	furi-uchi left	furi-uchi right	shita-uchi left	shita-uchi right	tsuki left	tsuki right
1	20	20	20	20	20	20	20	20	20	20	21	21	21	21	21	20
2	20	19	19	20	20	20	20	20	20	20	21	21	21	21	21	21
3	21	21	21	21	21	21	21	21	21	20	20	22	21	21	21	21
4	21	20	21	21	21	21	21	21	21	22	21	21	21	21	21	21
5	21	21	21	21	21	21	21	21	21	21	21	20	22	22	21	21
6	21	20	21	21	21	21	21	21	21	21	21	22	22	22	21	22

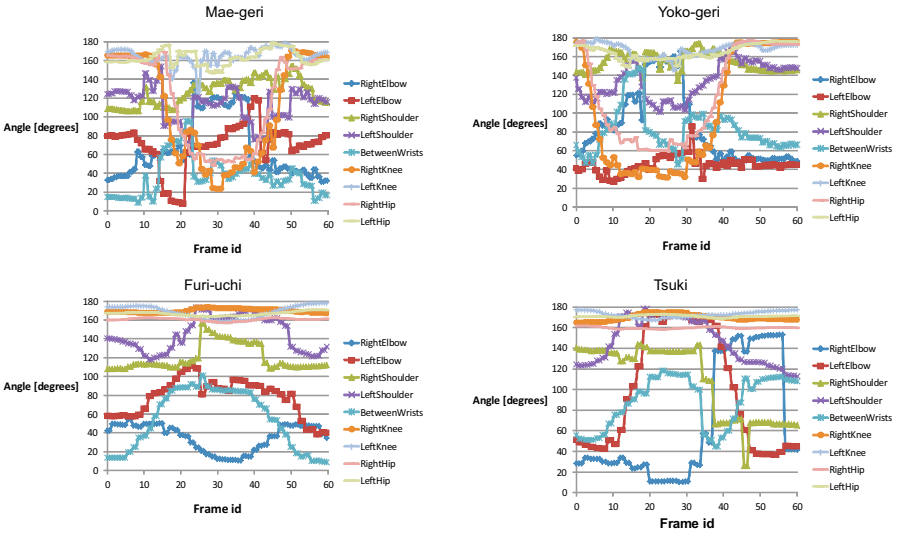

Fig. 4 This figure presents four plots of 9-dimensional angle-based signals of exemplar karate techniques from our datasets resampled to 60 frame length. Kicks were performed by *right leg* while punches with *left hand*

In Fig. 4 we present four plots of 9-dimensional angle-based signals of exemplar karate techniques from our datasets. We can clearly see that kicking actions highly involve whole body while punching mostly hands and marginally rest of the body that agrees with those movements motoric.

In both experiments we used features sets described in Sect. 2.1. We have implemented our solution in R language using H2O package H2O (2015) for NN and RF and kernlab Karatzoglou et al. (2004) for SVM.

In first experiment on gym exercises dataset (Table 1) we have used three types of classifiers. The first was NN with 4, 8, 16, 64, 128 and 256 neurons in hidden layer (see Sect. 2.2), the RF with 4, 8, 16, 64, 128 and 256 trees (see Sect. 2.3) and SVM (see Sect. 2.3). We have also compared obtained results with multivariate continuous hidden Markov model classifier with 4 hidden states from Hachaj and Ogiela (2015b) which also used angle-based and distance-based features set, however not the same as we proposed in Sect. 2.1. Table 3 presents averaged recognition rate (RR) for gym exercises obtained with k-fold cross-validation plus/minus standard deviation.

In Figs. 5 and 6 we present visualization of data from Table 3. Color bars represent averaged RR and black bars stand for standard deviation.

In second experiment on karate techniques dataset (Table 2) we have used the very similar classifiers settings as in previous one. We have also compared obtained results with multivariate continuous hidden Markov model classifier with 4 hidden states from Hachaj et al. (2015a) which used angle-based features set, however not the same as we proposed in Sect. 2.1. Also dataset in Hachaj et al. (2015a) did not contain six classes of actions namely punches. Table 4 presents averaged (RR) for karate techniques obtained with k-fold cross-validation plus/minus standard deviation.

In Figs. 7 and 8 we present visualization of data from Table 4. Color bars represent averaged RR and black bars stand for standard deviation.

Table 3 This table presents averaged recognition rate for gym exercises obtained with k-fold cross-validation plus/minus standard deviation. Values in table header are numbers of neurons in hidden layer of NN or numbers of trees in RF

Features	Classifier	4 (%)	8 (%)	16 (%)	32 (%)	64 (%)	128 (%)	256 (%)	Classifier	RR (%)
Angle-based	NN	95 ± 6	99 ± 2	99 ± 1	99 ± 1	98 ± 2	98 ± 3	100 ± 0	SVM	99 ± 3
Angle-based	RF	99 ± 1	97 ± 4	98 ± 2	99 ± 2	99 ± 1	99 ± 2	99 ± 1	HMM0	97 ± 14
Distance-based	NN	93 ± 6	99 ± 2	100 ± 0	99 ± 1	98 ± 1	98 ± 2	99 ± 1	SVM	97 ± 3
Distance-based	RF	93 ± 7	97 ± 2	99 ± 2	99 ± 2	99 ± 2	99 ± 2	99 ± 1	HMM0	95 ± 3

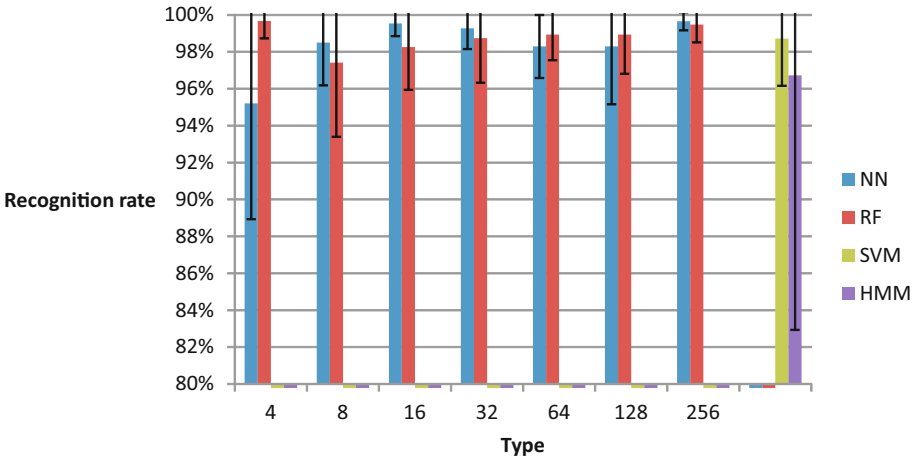

Fig. 5 This figure presents visualization of data from Table 3 on angle-based features. *Color bars* represent averaged RR and *black bars* stand for standard deviation. Values on *horizontal axis* are numbers of neurons in hidden layer of NN or numbers of trees in RF

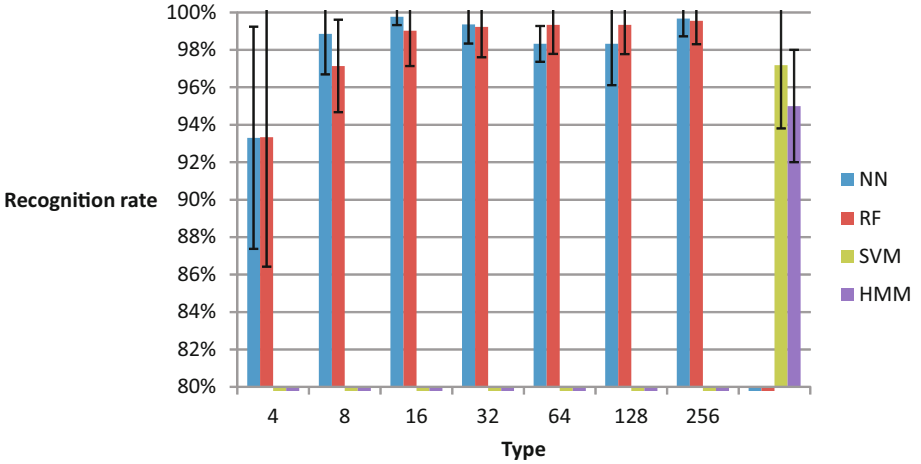

Fig. 6 This figure presents visualization of data from Table 3 on distance-based features. *Color bars* represent averaged RR and *black bars* stand for standard deviation. Values on *horizontal axis* are numbers of neurons in hidden layer of NN or numbers of trees in RF

4 Discussion

The results we obtained on gym exercises dataset are very promising. None of tested classifier's setup has RR below 93 %. Comparing to HMM Hachaj and Ogiela (2015b) the methodologies we have proposed in this paper have significantly lower variance. The high variance in HMM was caused by the fact that sll and slr were often confused with each other. That is caused by limitation of HMM model that cannot represent the necessary amount of movement trajectory without losing the generalization ability. In gym dataset there is not much difference in RR between used pattern classification techniques. It seems that in NN when number of neurons in hidden layer is greater or equal to 8, RF with at least 32 trees

Table 4 This table presents averaged recognition rate for karate techniques obtained with k-fold cross-validation plus/minus standard deviation. Values in table header are numbers of neurons in hidden layer of NN or numbers of trees in RF

Features	Classifier	4 (%)	8 (%)	16 (%)	32 (%)	64 (%)	128 (%)	256 (%)	Classifier	RR (%)
Angle-based	NN	59 ± 23	81 ± 14	93 ± 6	96 ± 5	97 ± 3	96 ± 4	97 ± 2	SVM	97 ± 3
Angle-based	RF	90 ± 9	94 ± 5	95 ± 4	97 ± 4	97 ± 3	97 ± 4	97 ± 3	HMM	97 ± 2
Distance-based	NN	60 ± 16	78 ± 13	91 ± 9	93 ± 7	93 ± 7	92 ± 7	91 ± 8	SVM	93 ± 6
Distance-based	RF	80 ± 17	87 ± 9	90 ± 10	92 ± 8	92 ± 8	94 ± 6	94 ± 6	–	–

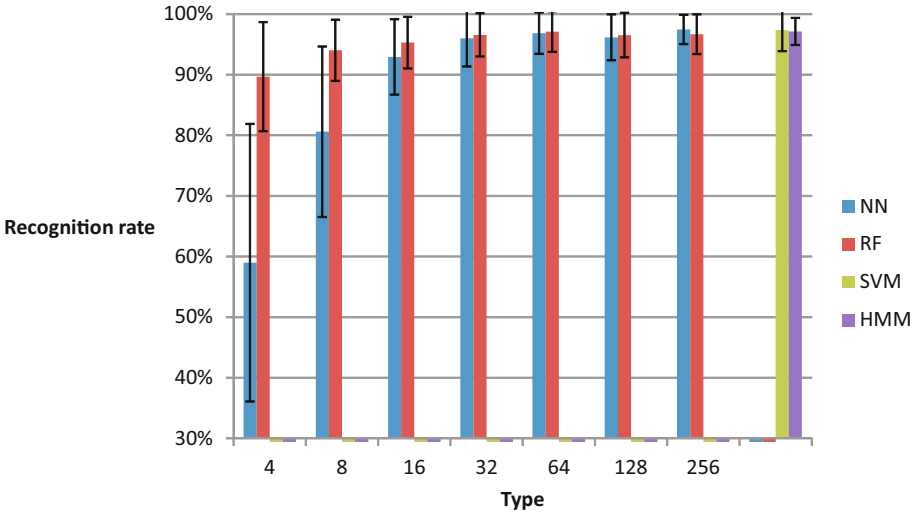

Fig. 7 This figure presents visualization of data from Table 4 on angle-based features. *Color bars* represent averaged RR and *black bars* stand for standard deviation. Values on *horizontal axis* are numbers of neurons in hidden layer of NN or numbers of trees in RF

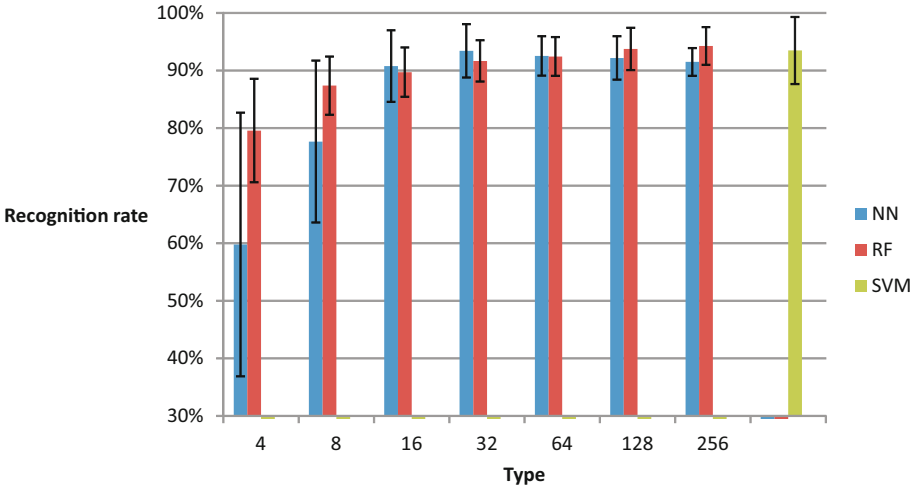

Fig. 8 This figure presents visualization of data from Table 4 on distance-based features. *Color bars* represent averaged RR and *black bars* stand for standard deviation. Values on *horizontal axis* are numbers of neurons in hidden layer of NN or numbers of trees in RF

and SVM with Gaussian radial basis kernel with angle-based features the RR reaches 99 % or even 100 %. We might conclude that applying all examined pattern recognition techniques (NN, RF and SVM) for both types of features representation resulted in equally very good classification results.

The karate techniques dataset is more difficult for correct recognition than previous one. It is because it has more classes of movements (16 comparing to 9 of gym's). None of the methods exceeded 97 % of RR. This time we can clearly observe that angle-based representation

gives better RR than distance-based one. Mostly often errors were caused by misclassification of punches (most notable furi-uchi with tsuki) and blocks (uchi-uke and gedan-barai and jodan-uke). This is caused by low quality of data tracking and heavy tracking errors that becomes visible when hands are positioned near other body parts. The angle-based features derived from joints positions seem to be more resistant to those noises than distance-based features. The highest RR ($97 \pm 2\,\%$) was obtained for NN with 256 neurons in hidden layer that uses angle-based coordinates. Those results were similar to HMM Hachaj et al. (2015a) which was also $97 \pm 2\,\%$, however we must notice that karate dataset from Hachaj et al. (2015a) did not include punches (6 additional classes of movements) and we might expect that finally RR of HMM will be far worse than 97 %. Also SVM classifier and RF with 64 and 256 trees have very similar RR namely 97 % with only slightly higher standard deviation ($\pm 3\,\%$ in SVD and in RF).

5 Conclusions

The proposed movement data representation technique based on resampling the input multi-dimensional signal to common length resulted in high RR to all applied pattern recognition methods. Basing on our experiments on relatively large datasets (9 classes with 770 actions samples and 16 classes with 1996 samples) it seems that angle-based 9-dimensioanl features set guaranteed higher RR than 15 or 16 distance-based features set. That is due the fact that angle based features seem to be more resistant to tracking inaccuracies present in the dataset. The most important aspect while choosing appropriate classifier is to select a method that is capable to operate on data sample with many dimensions (in our case between 540 and 960). This type of actions recognition approach outperforms key frame-based approach that uses multivariate continuous hidden Markov model classifier. Our method is also easy to setup and does not require many adaptive parameters to work successfully. Results presented in this work give very important hints for developing similar actions recognition systems because easy to repeat features selection and classification setup we have described seems to guarantee high efficiency and effectiveness of overall solution.

The goal for the future is to apply the proposed data representation schema for quantitative analysis of actions. The most straightforward but promising approach might be using NN with auto-encoding architecture which is effective approach in anomaly and outliners detection Candel and Parmer (2015). We believe that this type of analysis will be useful in outdoor real-time hazardous situation detection and high-quality body actions analysis (especially in sport).

Acknowledgements This work has been supported by the National Science Centre, Republic of Poland, under project number 2015/17/D/ST6/04051. We kindly thank company NatuMed Sp. z o.o (Targowa 17a, 42-244 Wancerzow, Poland) for supplying us with gym exercises SKL dataset that together with our own SKL recordings of gym and karate was used as training and validation dataset in this research.

References

Bilen, H., Namboodiri, V. P., & Van Gool, L. J. (2014). Object and action classification with latent window parameters. *International Journal of Computer Vision*, *106*(3), 237–251.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.
Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Monterey: Wadsworth & Brooks/Cole Advanced Books & Software.
Burghouts, G. J., & Schutte, K. (2013). Spatio-temporal layout of human actions for improved bag-of-words action detection. *Pattern Recognition Letters archive*, *34*(15), 1861–1869. doi:10.1016/j.patrec.2013.01.024.
Burghouts, G. J., Schutte, K., Bouma, H., & den Hollander, R. J. M. (2014). Selection of negative samples and two-stage combination of multiple features for action detection in thousands of videos. *Machine Vision and Applications*, *25*(1), 85–98.
Candel, A. & Parmer, V. (2015), Deep Learning with H2O, Published by H2O, 2015, http://leanpub.com/deeplearning (Access date August 08, 2015)
Cao, X., Zhang, H., Deng, C., Liu, Q., & Liu, H. (2014). Action recognition using 3D DAISY descriptor. *Machine Vision and Applications*, *25*(1), 159–171.
Charalampous, K., & Gasteratos, A. (2014). A tensor-based deep learning framework. *Image and Vision Computing*, *32*(11), 916–929.
Chen, G., Clarke, D., Giuliani, M., Gaschler, A., & Knoll, A. (2015). Combining unsupervised learning and discrimination for 3D action recognition. *Signal Processing*, *110*, 67–81.
Chen, W., & Guo, G. (2015). TriViews: A general framework to use 3D depth data effectively for action recognition. *Journal of Visual Communication and Image Representation*, *26*, 182–191.
Coleca, F., Klement, S., Martinetz, T. & Barth, E. (2013), Real-time skeleton tracking for embedded systems, *Proceedings SPIE* 8667, Multimedia Content and Mobile Devices, 86671X (March 7, 2013), doi:10.1117/12.2003004
del Rincón, J. M., Santofimia, M. J., & Nebel, J.-C. (2013). Common-sense reasoning for human action recognition. *Pattern Recognition Letters*, *34*(15), 1849–1860.
Díaz-Más, L., Muñoz-Salinas, R., Madrid-Cuevas, F. J., & Medina-Carnicer, R. (2012). Three-dimensional action recognition using volume integrals. *Pattern Analysis and Applications*, *15*(3), 289–298.
Fan, Y.-J., & Chaovalitwongse, W. A. (2010). Optimizing feature selection to improve medical diagnosis. *Annals of Operations Research*, *174*(1), 169–183. doi:10.1007/s10479-008-0506-z.
Guo, W., & Chen, G. (2015). Human action recognition via multi-task learning base on spatial-temporal feature. *Information Sciences*, *320*(1), 418–428.
H_2O (2015), Official website of H_2O machine learning programming library. http://h2o.ai/, Accessed August 08, 2015.
Hachaj, T., & Ogiela, M. R. (2014). Rule-based approach to recognizing human body poses and gestures in real time. *Multimedia Systems*, *20*(1), 81–99.
Hachaj, T., & Ogiela, M. R. (2015a). Full body movements recognition - unsupervised learning approach with heuristic R-GDL method. *Digital Signal Processing*, *46*, 239–252. doi:10.1016/j.dsp.2015.07.004.
Hachaj, T. & Ogiela, M. R. (2015b), Human actions recognition on multimedia hardware using angle-based and coordinate-based features and multivariate continuous hidden Markov model classifier, Multimedia Tools and Applications, In press, DOI:10.1007/s11042-015-2928-3
Hachaj, T., Ogielaj, M. R., & Koptyra, K. (2015a). Application of assistive computer vision methods to Oyama Karate techniques recognition. *Symmetry*, *7*(4), 1670–1698. doi:10.3390/sym7041670.
Hachaj, T., Ogiela, M.R. & Koptyra, K. (2015b). *Effectiveness comparison of Kinect and Kinect 2 for recognition of Oyama karate techniques, NBiS 2015 - The 18-th International Conference on Network-Based Information Systems (NBiS 2015)*, September 2–4, Taipei, Taiwan, pp. 332–337, DOI: 10.1109/NBiS.2015.51, ISBN: 978-1-4799-9942-2/15
Ji, X.-F., Wu, Q.-Q., Ju, Z.-J., & Wang, Y.-Y. (2014). Study of human action recognition based on improved spatio-temporal features. *International Journal of Automation and Computing*, *11*(5), 500–509.
Jiang, M., Kong, J., Bebis, G., & Huo, H. (2015). Informative joints based human action recognition using skeleton contexts. *Signal Processing: Image Communication*, *33*, 29–40.
Jiang, Z., Linb, Z., & Davis, L. S. (2012). Class consistent k-means: Application to face and action recognition. *Computer Vision and Image Understanding*, *116*(6), 730–741.
Jiang, Z., Linb, Z., & Davis, L. S. (2013). A unified tree-based framework for joint action localization, recognition and segmentation. *Computer Vision and Image Understanding*, *117*(10), 1345–1355.
Jiu, M., Wolf, C., Garcia, C., & Baskurt, A. (2012). Supervised learning and codebook optimization for bag-of-words models. *Cognitive Computation*, *4*(4), 409–419.
Karatzoglou A., Smola, A., Hornik, K. & Zeileis, A. (2004). Kernlab– An S4 Package for Kernel Methods in R, *Journal of Statistical Software,11* (9)
Knerr, S., Personnaz, L. & Dreyfus, G. (1990). *Single-layer learning revisited: A stepwise procedure for building and training a neural network, Neurocomputing, Volume 68 of the series NATO ASI Series* pp. 41–50.

Kreßel, U. H.-G. (1999). *Pairwise classification and support vector machines. Advances in kernel methods*. MA: MIT Press Cambridge.
Längkvist, M., Karlsson, L., & Loutfi, A. (2014). A review of unsupervised feature learning and deep learning for time-series modeling. *Pattern Recognition Letters*, *42*, 11–24.
LeCun, Y., Bottou, L., Orr, G. B. & Müller, K.R (2002). Neural networks: Tricks of the trade, *Volume 1524 of the series Lecture Notes in Computer Science*, pp. 9–50.
Li, N., Cheng, X., Zhang, S., & Wu, Z. (2014). Realistic human action recognition by Fast HOG3D and self-organization feature map. *Machine Vision and Applications*, *25*(7), 1793–1812.
Li, S., Liu, Z.-Q., & Chan, A. B. (2015). Heterogeneous multi-task learning for human pose estimation with deep convolutional neural network. *International Journal of Computer Vision*, *113*(1), 19–36.
Liu, C. W., Pei, M. T., Wu, X. X., Kong, Y., & Jia, Y. D. (2014). Learning a discriminative mid-level feature for action recognition. *Science China Information Sciences*, *57*(5), 1–13.
Liu, A.-A., Nie, W.-Z., Su, Y.-T., Ma, L., Hao, T., & Yang, Z.-X. (2015). Coupled hidden conditional random fields for RGB-D human action recognition. *Signal Processing*, *112*, 74–82.
Liu, L., Shao, L., & Rockett, P. (2013a). Boosted key-frame selection and correlated pyramidal motion-feature representation for human action recognition. *Pattern Recognition*, *46*(7), 1810–1818.
Liu, L., Shao, L., & Rockett, P. (2013b). Human action recognition based on boosted feature selection and naive Bayes nearest-neighbor classification. *Signal Processing*, *93*(6), 1521–1530.
Mahbub, U., Imtiaz, H., & Ahad, A. R. (2014). Action recognition based on statistical analysis from clustered flow vectors. *Signal, Image and Video Processing*, *8*(2), 243–253.
Nasiri, J. A., Charkari, N. M., & Mozafari, K. (2014). Energy-based model of least squares twin Support Vector Machines for human action recognition. *Signal Processing*, *104*, 248–257.
Omidyeganeh, M., Ghaemmaghami, S., & Shirmohammadi, S. (2013). Application of 3D-wavelet statistics to video analysis. *Multimedia Tools and Applications*, *65*(3), 441–465.
Papadopoulos, G. T., Axenopoulos, A. & Daras, P., Real-time Skeleton-tracking-based Human Action Recognition Using Kinect Data (2014), MultiMedia Modeling, Volume 8325 of the series Lecture Notes in Computer Science pp. 473–483
Pazhoumand-Dar, H., Lam, C.-P., & Masek, M. (2015). Joint movement similarities for robust 3D action recognition using skeletal data. *Journal of Visual Communication and Image Representation*, *30*, 10–21.
Recht, B., Re, C., Wright, S. & Niu, F. (2011). Hogwild: A Lock-Free Approach to Parallelizing Stochastic Gradient Descent, Advances in Neural Information Processing Systems 24, Editor J. Shawe-taylor and R.s. Zemel and P. Bartlett and F.c.n. Pereira and K.q. Weinberger, pp. 693–701, 2011
Saito, Y., & Nishiyama, H. (2015). Design of a collaborative method with specified body regions for activity recognition: generating a divided histogram considering occlusion. *Artificial Life and Robotics*, *20*(2), 129–136.
Schölkopf, B., & Smola, A. J. (2002). *Learning with kernels*. Cambridge: MIT Press.
Shen, H., Yan, Y., Xu, S., Ballas, N., & Chen, W. (2015). 2015. *Evaluation of semi-supervised learning method on action recognition, Multimed Tools Appl*, *74*, 523–542. doi:10.1007/s11042-014-1936-z.
Shotton, J., Sharp, T., Kipman, A., Fitzgibbon, A., Finocchio, M., Blake, A., Cook, M. & Moore, R. (2013), Real-time human pose recognition in parts from single depth images, Communications of the ACM, Volume 56 Issue 1, January 2013, pp. 116–124, ACM New York, NY, USA.
Vapnik, V. N. (1998). *Statistical learning theory*. New York: Wiley.
Vrigkas, M., Karavasilis, V., Nikou, C., & Kakadiaris, I. A. (2014). Matching mixtures of curves for human action recognition. *Computer Vision and Image Understanding*, *119*, 27–40.
Wu, J., Hu, D., & Chen, F. (2014). Action recognition by hidden temporal models. *The Visual Computer*, *30*(12), 1395–1404.
Yahav, I., & Shmueli, G. (2014). Outcomes matter: estimating pre-transplant survival rates of kidney-transplant patients using simulator-based propensity scores. *Annals of Operations Research*, *216*(1), 101–128. doi:10.1007/s10479-013-1359-7.
Yang, X., & Tian, Y. (2014). Effective 3D action recognition using EigenJoints. *Journal of Visual Communication and Image Representation*, *25*(1), 2–11.
Zhen, X., Shao, L., & Li, X. (2014). Action recognition by spatio-temporal oriented energies. *Information Sciences*, *281*, 295–309.
Zhu, F., Shao, L., & Lin, M. (2013). 2013. *Multi-view action recognition using local similarity random forests and sensor fusion, Pattern Recognition Letters*, *34*(1), 20–24.
Ziaeefard, M., & Bergevin, R. (2015). 2015. *Semantic human activity recognition: A literature review, Pattern Recognition*, *48*(8), 2329–2345.