Multi-Sensor Acceleration-based Action Recognition*

Florian Baumann[1], Irina Schulz[2], Bodo Rosenhahn[1]

[1] Institut für Informationsverarbeitung (TNT), Leibniz Universität Hannover, Germany
[2] Institute for Systems Engineering (RTS), Leibniz Universität Hannover, Germany

Abstract. In this paper, a framework to recognize human actions from acceleration data is proposed. An important step for an accurate recognition is the pre-processing of input data and the following classification by the machine learning algorithm. In this paper, we suggest to combine Dynamic Time Warping (DTW) with Random Forest. The intention of using DTW is to pre-process the data to eliminate outliers and to align the time series. Many applications require more than one inertial sensor for an accurate prediction of actions. In this paper, nine inertial sensors are deployed to ensure an accurate recognition of actions. Further, sensor fusion approaches are introduced and the most promising strategy is shown. The proposed framework is evaluated on a self-recorded dataset consisting of six human actions. Each action was performed three times by 20 subjects. The dataset is publicly available for download.

1 Introduction

In recent years, the use of inertial sensors has become a popular topic in machine learning. One reason for the growing interest is the improved quality and the reduced costs of the hardware [9]. Another reason is the rising number of applications. For instance, sonification of movements [8, 10], analysis of sports-, rehabilitation-, and healthcare sessions [14, 21, 22, 32] as well as applications within the clinical and veterinary field [18, 24]. These applications require an accurate and precise recognition of actions and movements leading to a challenging topic in machine learning. For instance, each actor has the own style of performing an action and many variations in the subject's movement are possible. Thus, a large intra-class variation is inevitable. These problems are also reflected in the recorded data: the gathered acceleration-based time series strongly differ in their amplitude and length.

Contribution In this paper, a combination of Dynamic Time Warping (DTW) with the well-known machine learning algorithm Random Forest is proposed. DTW is used as a pre-processing step to eliminate outliers, to align different time series and to compensate large intra-class variations. For classification, a Random Forest is learned on the aligned, raw acceleration values.

The proposed approach is applied to a self-recorded dataset. Inspired by the KTH dataset for single human action recognition [29], six actions were defined: *walking, running, jogging, boxing, clapping and waving*. Each action was performed three times by 20 subjects. The acceleration data was gathered by nine Xsens motion wireless tracker

* This work has been partially funded by the ERC within the starting grant Dynamic MinVIP.

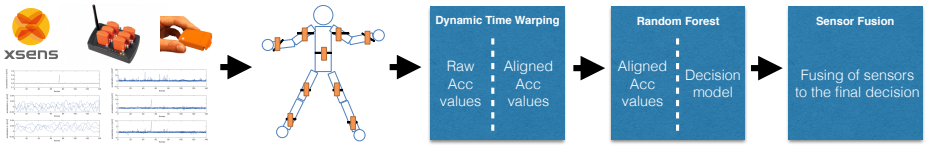


Fig. 1. An overview about the proposed framework. Acceleration data of six actions are obtained by nine inertial sensors. A Dynamic Time Warping is used in a pre-processing step to prepare the input data. The aligned acceleration time series are directly used to learn a Random Forest classifier. The final decision is determined by using sensor fusion methods.

(MTw). Each subject was equipped with sensors on the left/right wrist, left/right upper arm, left/right thigh, left/right ankle and one MTw on the waist, also see Figure 1. To combine multiple sensors, fusion methods are presented and evaluated. The dataset is publicly available for download. Thus, other researchers can evaluate their methods and algorithms and publish competing results.
The paper is structured as follows. Section 2 gives a short overview about related work. Section 3 briefly describes Dynamic Time Warping, Random Forest and sensor fusion strategies. Section 4 presents the dataset and the experimental results. Section 5 concludes the paper and gives an overview about future work.

2 Related Work

Action recognition has been playing an important role in many areas of medicine, in the industrial domain, in the automotive area, for scene understanding or in the surveillance area [1, 25]. Many works use acceleration-based sensors, referred to as Xsens-sensors for ambulatory measurement [23] or for clinical gait analysis [12]. These works reveal that physical activities can be well-recognized by inertial sensors.
Chambers et al. [11] used one inertial sensor that was attached to the wrist of a subject for complex gesture recording by using Hidden Markov Models. Wang et al. [34] attached sensors to subjects and learned a Support Vector Machine to recognize daily activities. Karantonis et al. [17] explained a basic decision tree method using a single sensor on the waist of subjects for real-time classification. Tautges et al. [33] reconstructed whole body motions from the data taken by as few as one, two and four inertial sensors for several classes of motions. They use the acceleration data from some Xsens inertial sensors attached to the hands and feet of some subjects to reconstruct the performed motions. These motions are compared to a video that was taken during the capturing. Further information and a detailed survey about the current research using inertial sensors is presented by Avci et al. [3].
In comparison to the above mentioned works, a framework based on Random Forest in combination with Dynamic Time Warping is proposed. Six typical actions are defined and a self-recorded dataset is provided to the community. By using this dataset, nine inertial sensors can be utilized to recognize and analyze human actions. Finally, the most promising sensor fusion strategy is presented.

3 Approach

Figure 1 presents an overview of the proposed framework. The acceleration time series in (x, y, z)-direction are gathered by nine inertial sensors. For aligning the time series to each other a Dynamic Time Warping is applied to every training and testing example. Instead of deploying a specific feature extraction method, the aligned (x, y, z)-acceleration values are directly used to learn a Random Forest classifier. The final decision is computed by a sensor fusion method.
This Section briefly describes Dynamic Time Warping for pre-processing, Random Forest for classification and the sensor fusion strategies.

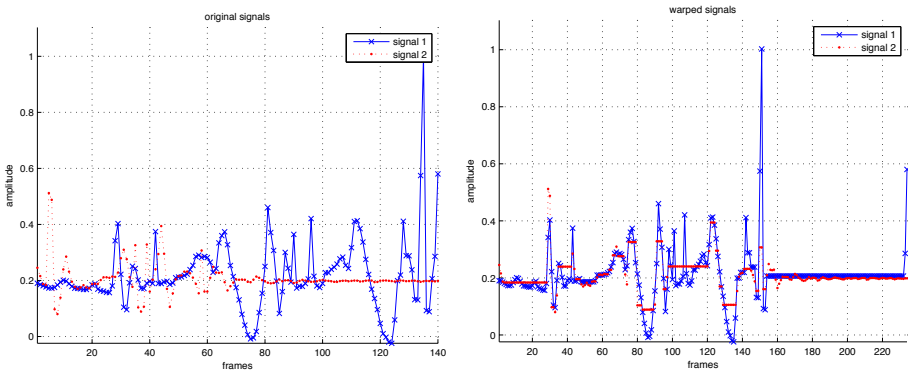
3.1 Dynamic Time Warping

Dynamic Time Warping (DTW) was introduced by Bellman and Kalaba [4]. DTW has been applied to several fields of applications like video or audio data for measuring the similarity of two temporal sequences. For instance, Myers et al. and Sakoe et al. applied DTW to the task of speech recognition [20, 27]. Generally, DTW is an algorithm for mapping values between two temporal sequences to each other.
In the following, a brief explanation of the theory is given. First, the algorithm applies a distance between any two values of the signals using a weighting function, such as the euclidean distance for each parameter of each tuple. The output is referred to as a cost function. In the next step the algorithm seeks the lowest cost from the start to the end of both signals over the stretched matrix of pairwise current cost of all points of both signals. The actual path, referred to as a warping, is determined by backtracking the first pass of the algorithm. The backtracking allows a precise representation of each point of the shorter signal to one or more points of the longer signal. Thus the approximate time distortion is represented. Further information and a detailed review are presented by Senin [30].
Figure 2(a) illustrates two acceleration signals in x-direction of a boxing gesture. The signals differ in their amplitude. Signal 1 is defined as reference signal. Figure 2(b) presents the warping of signal 2 to the reference signal 1. The amplitudes are nearly the same and signal 2 is aligned.
In this work the standard DTW algorithm with a time complexity of $\mathcal{O}(N^2)$ is implemented. For a real-time capable modification Rakthanmanon et al. [26] propose a combination of four approaches to search and mine time series in a very efficient way.
After applying the DTW to each training and testing example a Random Forest is used to find discriminative (x, y, z)-acceleration values.

3.2 Random Forest

Random Forest was published by Leo Breiman in 2001 [6]. It is a substantial modification of bagging [5] with a random feature selection proposed by Ho [15, 16] and Amit [2]. A Random Forest consists of a collection of CART-like (Classification and Regression Tree) decision trees h_t, $1 \leq t \leq T$, [7]:

$$\{h(\boldsymbol{x}, \Theta_t)_{t=1,\ldots T}\}$$



(a) (b)

Fig. 2. (a) Original signals of the x-acceleration of a boxing action. The signals differ in their amplitude. (b) Acceleration signal in x-direction of a boxing action. Signal 2 is warped to signal 1. The amplitudes are nearly the same and outliers are compensated.

where $\{\Theta_t\}$ is a bootstrap sample from the training data. Each tree casts a vote on a class for the input $\boldsymbol{x}$. The class probabilities are estimated by majority voting and used to calculate the sample's label $y(\boldsymbol{x})$ with respect to a given feature vector $\boldsymbol{x}$:

$$y(\boldsymbol{x}) = \underset{c}{\operatorname{argmax}} \left(\frac{1}{T} \sum_{t=1}^{T} F_{h_t(\boldsymbol{x})=c} \right) \tag{1}$$

The decision function $h_t(\boldsymbol{x})$ returns the resulting class c of one tree with the indicator function F:

$$F_{h_t(\boldsymbol{x})=c} = \begin{cases} 1, & h_t(\boldsymbol{x}) = c, \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

Random Forest has a high classification accuracy and can deal with large data sets for multiple classes with outstanding time efficiency [6].

Classification Time-series are classified by passing them down each tree until a leaf node is reached. The resulting class is defined by each leaf node and the final decision is determined by taking the class having the most votes (majority vote), see Equation (1).

3.3 Sensor Fusion

The sensor fusion part describes methods to combine decisions by different classifiers (or sensors) to the final decision. In this paper, the sensor fusion part has to handle the information of nine sensors. A feature is represented by a single DTW aligned acceleration value.

For all experiments, the input data for the Random Forest is composed of the concatenated acceleration values (a_x, a_y, a_z), each with m samples:

$$RF_{input} = [a_x(1), ..., a_x(m), a_y(1), ..., a_y(m), a_z(1), ..., a_z(m)], \tag{3}$$

RF_{input} is used for two approaches of determining the final decision:

1. Learning a Random Forest using all sensors (fusion is not required)
2. Learning a Random Forest for each sensor individually (fusion is required)

In the first experiment the input data for the Random Forest is composed by concatenating the acceleration values of n sensors:

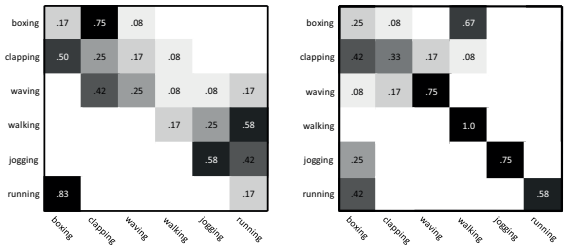
$$RF_{input_{All}} = [RF_{input_1}, ..., RF_{input_n}]. \tag{4}$$

This approach leads to a huge pool of possible values. A feature vector consists of $n \times 3 \times m$ values. A sensor fusion is not necessary, because the Random Forest is learned with the input of all sensors. Usually the Random Forest algorithm is able to handle large dimensions of training data [6] but due to the random feature selection mechanism, we assume that this approach leads to poor results. The probability of selecting a discriminative feature is lower due to the number of chosen variables from the feature vector: For each tree $v = \sqrt{p}$ variables[3] with p = $n \times 3 \times m$ are selected to build the tree.
For the second approach, a Random Forest is learned from each sensor individually. Thus, n sensors require n classifiers. Similar to the first approach $v = \sqrt{p}$ variables with $p = 3 \times m$ are selected to build the tree leading to $n \times \sqrt{3 \times m}$ variables. In comparison to the first approach the number of variables to split is three times higher. We assume that this approach reaches higher and more robust accuracies. But an additional step of fusing n probability distributions to the final decision is required.
Fusing: By taking the probability distributions of $n = 9$ sensors into consideration the following combination strategies for finally determining the decision are proposed:

1. Choose class with highest probability
2. Choose class with majority voting
3. Fusion of probability distributions by product law
4. Fusion of probability distributions by summation rule

For the first case, we assume that the most reliable class gains the highest probability. For the second case the class which gains the most votes of all classifiers is chosen. This idea is inspired by the majority tree voting of a Random Forest. For the third case the probability distributions of all classifiers are taken into consideration and fused by the product law. The final decision is determined by combining the class probabilities $\Pr(A_i) \cdots \Pr(I_i)$ of one class i from each sensor $A_i \cdots I_i$ with the product law[4]. For the fourth case the class probabilities are fused using the summation rule. The class probabilities $\Pr(A_i) \cdots \Pr(I_i)$ are summed up and the final decision is determined using the class with the highest probability.

[3] Random Forest is not restricted to use $v = \sqrt{p}$ variables. The number of variables can be freely chosen. Best results were obtained by taking $m = \sqrt{p}$ variables, as proposed by Breiman.
[4] With the assumption that sensors $A_i \cdots I_i$ are independent.



(a) (b)

Fig. 3. (a) Take the class with the highest probability by using all nine classifiers with an averaged accuracy of 26.50%. (b) Choose class with majority voting by using all sensors, the averaged accuracy is 61.00%.

4 Experimental Results

This Section describes the selection and recording of acceleration data and the classification experiments. First, the dataset is defined and presented followed by the sensor fusion strategies. Finally a detailed discussion about the examined experiments is given.

4.1 Self-recorded dataset

Inspired by the well-known KTH dataset for single human action recognition [29], six actions were defined: *walking, running, jogging, boxing, clapping and waving*. Each action was repeated three times by 20 subjects, leading to an overall dataset of 360 time series. The dataset is publicly available for download[5] in a Matlab file format.
The Xsens MTw Development Kit[6] was used for data recording. MTw stands for Motion Tracker wireless. It is a measuring instrument with built in 3D accelerometer, gyroscope, magnetometer (compass 3D) and a barometer (pressure sensor). The inertial sensors gather the acceleration values in the three-dimensional space. Each sensor determines the acceleration along (x, y, z)-axes. Figure 1 gives an overview of a subject equipped with nine sensors attached to the body.

4.2 Experiments

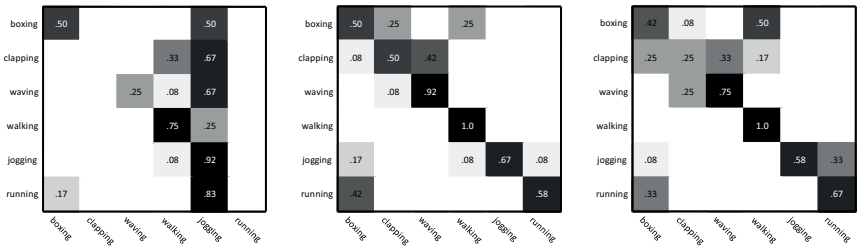
In this Section the proposed sensor fusion strategies are compared to each other. As mentioned in Section 3.3, we compare the strategy of learning a descriptor using all sensors to the strategy of learning a descriptor using each sensor individually. The optimal parameters for the Random Forest were determined by a cross validation mechanism and set to *maximum depth of a tree* $= 64$, *optimal number of trees* $= 8$ and *minimum number of leaves* $= 2$ for all experiments. The entropy was used as the splitting criterion. More information about different strategies are found in the literature [6]. As

[5] `http://www.tnt.uni-hannover.de/staff/baumann/`
[6] `http://www.xsens.com`

already discussed, using all sensors to learn a single Random Forest classifier results in poor accuracies. The classifier reached an averaged accuracy of 49.80%. The reason for these poor accuracies results from the lower number of variables to split each tree.
In the following experiments, the focus is on the evaluation of learning a classifier using each sensor individually. Results are reported for each fusion strategy individually:
Choose class with highest probability: The final decision is determined by taking the class with the highest probability of all nine classifiers.
Figure 3a presents the confusion matrix. The accuracy of 26.50% is quite low. The results are not unusual, since only one decision by one sensor was taken into account. This leads to a higher sensitivity to noise and outliers.
Choose class by majority voting of all classifiers: For this case, the final decision is determined by a majority voting of all sensors. Each Random Forest votes for one class while the majority class is chosen.
Figure 3b presents the confusion matrix for this experiment. The averaged accuracy is 61.00%. Since all decisions are fused by using a majority voting, the results are better leading to a more robust recognition. Walking is perfectly classified and waving and jogging reach 75.00% accuracy. Most confusions occur between boxing and clapping.
Fusion by product law: The probability distributions gained by all sensors are fused using the product law. For this experiment, three cases are compared:

1. Probabilities of 0% are taken into consideration
2. Probabilities of 0% are ignored
3. Introducing a threshold for taking only reliable probabilities

Figure 4a presents the results for the first case achieving an averaged accuracy of 42.00%. It is striking that most confusions occur between jogging and the other actions. Thus, taking all probabilities into account leads to low results.
Figure 4b presents a confusion matrix for the case of ignoring probabilities with 0% achieving the best accuracy of 70.00%. Walking is perfectly classified and waving reaches 92.00% accuracy. Boxing and clapping gain accuracies of 50% while most confusions occur between similar actions like boxing, waving and clapping.
Figure 4c presents the case of introducing a threshold. Only probabilities of more than



(a) (b) (c)

Fig. 4. Confusion matrices using the product law. (a) Taking all probabilities into consideration. Averaged accuracy is 42.00%. (b) Ignore poor decisions. Accuracy is 70.00%. (c) Thresholding the probabilities to take reliable decisions into consideration. The averaged accuracy is 61.00%.

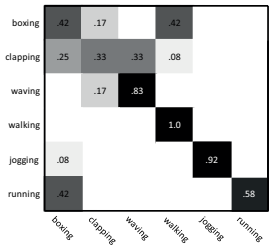


Fig. 5. Choose class using the summation rule. The averaged recognition rate is 68.00%.

40.00% are taken into consideration leading to an averaged accuracy of 61.00%. The threshold was empirically chosen. Also, for this experiment walking is perfectly classified and most confusions occur between boxing and clapping.

Fusion by summation rule: The probability distributions gained by all sensors are fused using the summation rule.

Figure 5 presents the confusion matrix for this case. The averaged accuracy is 68.00%. Walking is perfectly classified and waving and jogging reach high accuracies too. Also, for this experiment most confusions occur between boxing and clapping.

4.3 Discussion

In this Section two sensor fusion strategies were proposed. The first experiment of learning a Random Forest using all sensors results in low accuracies. The second method describes several fusion methods. Best results were achieved by learning a Random Forest for each sensor individually. Fusing the decisions by the product law leads to an accuracy of 70.00% and by using the summation rule to 68.00%. Walking is perfectly classified for nearly all experiments while most confusions occur between boxing and clapping.

5 Conclusions and Future Work

In this paper, nine inertial sensors were used to recognize six typical human actions: *walking, running, jogging, boxing, clapping and waving*. In a pre-processing step a Dynamic Time Warping is applied to align the acceleration time series to each other. The aligned time series are directly used to learn a Random Forest classifier. Furthermore, two sensor fusion strategies are proposed. By applying the product law or summation rule for fusing the class probabilities, accuracies up to 70.00% were reached. The self-recorded dataset is publicly available for download at `http://www.tnt.uni-hannover.de/staff/baumann/`.

Future Work Our plans for future work are to combine all decisions using the Dempster Shafer theory [31] or related works from the computer vision community basing on Dempster's theory of evidence, like [19, 28, 35]. Presumably, the results might be further improved.

Another idea is to realize a time series forest for classification and feature extraction

[13] and to spend more attention on the type of feature instead of using the raw acceleration data. It should also be made some experiments whether it is more convenient to use the orientations (quaternions) of each sensor instead of using only the acceleration data.

References

1. Aggarwal, J., Ryoo, M.: Human activity analysis: A review. ACM Computing Surveys 43(3), 16:1–16:43 (Apr 2011)
2. Amit, Y., Geman, D.: Shape quantization and recognition with randomized trees. Neural computation 9(7), 1545–1588 (1997)
3. Avci, A., Bosch, S., Marin-Perianu, M., Marin-Perianu, R., Havinga, P.: Activity recognition using inertial sensing for healthcare, wellbeing and sports applications: A survey. In: Architecture of Computing Systems (ARCS), 2010 23rd International Conference on (2010)
4. Bellman, R., Kalaba, R.: On adaptive control processes. Automatic Control, IRE Transactions on 4(2), 1–9 (1959)
5. Breiman, L.: Bagging predictors. Machine Learning 24, 123–140 (1996)
6. Breiman, L.: Random forests. Machine learning 45(1), 5–32 (2001)
7. Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J.: Classification and Regression Trees. Chapman & Hall, New York, NY (1984)
8. Brock, H., Schmitz, G., Baumann, J., Effenberg, A.O.: If motion sounds: Movement sonification based on inertial sensor data. In: 9th Conference of the International Sports Engineering Association (ISEA). Elsevier (Jan 2012)
9. Brückner, H.P., Nowosielski, R., Kluge, H., Blume, H.: Mobile and wireless inertial sensor platform for motion capturing in stroke rehabilitation sessions. In: Advances in Sensors and Interfaces (IWASI), 2013 5th IEEE International Workshop on. pp. 14–19 (2013)
10. Brückner, H.P., Wielage, M., Blume, H.: Intuitive and interactive movement sonification on a heterogeneous risc/dsp platform. The 18th Annual International Conference on Auditory Display (2012)
11. Chambers, G., Venkatesh, S., West, G., Bui, H.: Hierarchical recognition of intentional human gestures for sports video annotation. In: Pattern Recognition, 2002. Proceedings. 16th International Conference on (2002)
12. Cutti, A., Ferrari, A., Garofalo, P., Raggi, M., Cappello, A., Ferrari, A.: 'outwalk': a protocol for clinical gait analysis based on inertial and magnetic sensors. Medical and Biological Engineering and Computing 48(1), 17–25 (2010)
13. Deng, H., Runger, G., E. Tuv, M.V.: A time series forest for classification and feature extraction. Information Sciences 239(pp. 142-153.) (2013)
14. Ha, T.H., Saber-Sheikh, K., Moore, A.P., Jones, M.P.: Measurement of lumbar spine range of movement and coupled motion using inertial sensors–a protocol validity study. Manual Therapy 18(1), 87–91 (2013)
15. Ho, T.K.: Random decision forests. In: Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on. IEEE (1995)
16. Ho, T.K.: The random subspace method for constructing decision forests. Pattern Analysis and Machine Intelligence, IEEE Transactions on 20(8), 832–844 (1998)
17. Karantonis, D., Narayanan, M., Mathie, M., Lovell, N., Celler, B.: Implementation of a real-time human movement classifier using a triaxial accelerometer for ambulatory monitoring. Information Technology in Biomedicine, IEEE Transactions on 10(1), 156–167 (2006)
18. Lebel, K., Boissy, P., Hamel, M., Duval, C.: Inertial measures of motion for clinical biomechanics: Comparative assessment of accuracy under controlled conditions - effect of velocity. PLoS ONE 8(11) (2013)

19. Murphy, R.R.: Dempster-shafer theory for sensor fusion in autonomous mobile robots. Robotics and Automation, IEEE Transactions on 14(2), 197–206 (1998)
20. Myers, C., Rabiner, L., Rosenberg, A.: Performance tradeoffs in dynamic time warping algorithms for isolated word recognition. Acoustics, Speech and Signal Processing, IEEE Transactions on 28(6), 623–635 (1980)
21. van den Noort, J.C., Ferrari, A., Cutti, A.G., Becher, J.G., Harlaar, J.: Gait analysis in children with cerebral palsy via inertial and magnetic sensors. Medical & biological engineering & computing pp. 1–10 (2013)
22. Olsen, E., Haubro Andersen, P., Pfau, T.: Accuracy and precision of equine gait event detection during walking with limb and trunk mounted inertial sensors. Sensors (2012)
23. Parel, I., Cutti, A., Fiumana, G., Porcellini, G., Verni, G., Accardo, A.: Ambulatory measurement of the scapulohumeral rhythm: Intra- and inter-operator agreement of a protocol based on inertial and magnetic sensors. Gait and Posture 35(4), 636 – 640 (2012)
24. Pfau, T., Starke, S.D., Tröster, S., Roepstorff, L.: Estimation of vertical tuber coxae movement in the horse from a single inertial measurement unit. The Veterinary Journal (2013)
25. Poppe, R.: A survey on vision-based human action recognition. Image and Vision Computing 28(6), 976 – 990 (2010)
26. Rakthanmanon, T., Campana, B., Mueen, A., Batista, G., Westover, B., Zhu, Q., Zakaria, J., Keogh, E.: Searching and mining trillions of time series subsequences under dynamic time warping. In: Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 262–270. ACM (2012)
27. Sakoe, H., Chiba, S.: Dynamic programming algorithm optimization for spoken word recognition. Acoustics, Speech and Signal Processing, IEEE Transactions on 26(1), 43–49 (1978)
28. Scheuermann, B., Schlosser, M., Rosenhahn, B.: Efficient pixel-grouping based on dempsters theory of evidence for image segmentation. In: Lee, K., Matsushita, Y., Rehg, J., Hu, Z. (eds.) Computer Vision, ACCV 2012, Lecture Notes in Computer Science, vol. 7724, pp. 745–759. Springer Berlin Heidelberg (2013)
29. Schuldt, C., Laptev, I., Caputo, B.: Recognizing human actions: a local svm approach. In: Pattern Recognition. (ICPR). Proceedings of the 17th International Conference on (2004)
30. Senin, P.: Dynamic time warping algorithm review. Honolulu, USA (2008)
31. Shafer, G.: A mathematical theory of evidence, vol. 1. Princeton university press Princeton (1976)
32. Starrs, P., Chohan, A., Fewtrell, D., Richards, J., Selfe, J.: Biomechanical differences between experienced and inexperienced wheelchair users during sport. Prosthetics and Orthotics International 36(3), 324–331 (2012)
33. Tautges, J., Krüger, B., Zinke, A., Weber, A.: Reconstruction of human motions using few sensors
34. Wang, S., Yang, J., Chen, N., Chen, X., Zhang, Q.: Human activity recognition with user-free accelerometers in the sensor networks. In: Neural Networks and Brain, 2005. ICNN B. International Conference on (2005)
35. Wu, H., Siegel, M., Stiefelhagen, R., Yang, J.: Sensor fusion using dempster-shafer theory [for context-aware hci]. In: Instrumentation and Measurement Technology Conference, 2002. IMTC/2002. Proceedings of the 19th IEEE. vol. 1, pp. 7–12. IEEE (2002)