Human Action Recognition Using Laban Movement Analysis and Dynamic Time Warping

Zahra Ramezanpanah, Malik Mallem, Frédéric Davesne

Human Action Recognition Using Laban Movement Analysis and Dynamic Time Warping

Zahra Ramezanpanah[a], Malik Mallem[a], Frederic Davesne[a]

[a]*IBISC, University Evry Val d'Essonne, University Paris Saclay, 91000, Evry, France*

Abstract

Bilateral interaction between humans and robots is one of the areas that has attracted much attention in recent years. Automation of human behavior recognition is one of the main steps in achieving this goal. In this regard, in this paper we have designed a process for automatic identification of human gestures. The process consists of two main parts. In the first part, by using the Laban Movement Analysis method, we define a robust descriptor, and in the second part, we determine the robustness of the descriptor using the Dynamic time Warping algorithm. The method proposed in this paper has been tested on four public data-sets namely MSR Action 3D, Florence 3D actions, UTKinect-Action3D and SYSU 3D HUMAN-OBJECT INTERACTION data-sets. Given the results obtained from previous work, the efficiency of the proposed method can be more accurately understood. The results obtained confirm the effectiveness and the performance of our model which outperforms results presented in similar works on action recognition.

Keywords: Laban Movement Analysis; Dynamic Time Warping; Human Gesture Recognition.

1. Related works

Due to the complexity of software and hardware in working with computers, researchers have done great studies in the recent years to create an easy interaction with computers. One of the goals of Human computer interaction (HCI) is to enable computers to interact with humans using their voice, face, and gestures, instead of using computer hardware and programming techniques. Understanding human needs through analyzing audio and video data coming from sensors is the most important part of HCI. Human Action Recognition technology, one of the most important and widely used subdivisions of machine learning, mostly uses the concept of pattern recognition techniques. The gesture recognition process

* Corresponding author. Tel.: +0-000-000-0000 ; fax: +0-000-000-0000.
E-mail address: zahra.ramezanpanah@univ-evry.fr; malik.mallem@univ-evry.fr; frederic.davesne@univ-evry.fr

consists of two steps: the first stage is the data acquisition, feature extraction and their characterization, and the second is the classification of the extracted features based on the pattern formed in the first step. In this stage, the computer obtains its data from various sensors, which can be either two-dimensional or 3D cameras [24], or can be inertial sensors [40]. Also, in some cases, it can be a computer mouse or keyboard [18]. In the following, we will give an overview of research conducted in this field.

In the recent year, gesture recognition based on the 3D skeletal data, due to the ease of extracting data, has attracted much attention in multi media applications. For example in [26], the authors modeled a three-dimensional block whose dimensions are, respectively, the number of skeleton joints, the number of consecutive frame, and the three spatial coordinates (x, y, z) of the joint. They combined the Convolutional Neural Network (CNN) and the Long Short-Term Memory (LSTM) recurrent network as a machine learning method to recognize human action and hand gesture. In another study, [38], the researchers reshaped the 3D spatio-temporal data into three texture 2D images through color encoding, Joint Trajectory Maps (JTMs), and implemented Convolutional Neural Network to train the discriminative features for classifying human gestures. Also in [33], based on the 3D coordinates of the joints, a graph-based structure is proposed for gesture recognition. In this work, the joints and their dependency were considered as a graph. Vertices of the graph contain the 3D coordinates of the body joints, while the adjacency matrix captures their relationship.

Since the use of Deep Learning requires a lot of data and large data sets, in cases where not much data is available, classical methods are also widely used in this field. For example in [10], the authors used the joint angles and orientations of the most informative body parts to define their descriptor. Because they evaluated the proposed descriptor in small-sized data-sets, they used Support Vector Machine (SVM) for training and classification. The author in [22], defined their descriptor by pairwise relative positions of skeleton joints. Using the latent SVM method, they demonstrated the effectiveness of their proposed method on three datasets called MSR Action 3D, Florence 3D actions and UTKinect-Action3D. Other classical methods of classification in the field of pattern recognition that have attracted the attention of researchers include Hidden Makov Model (HMM), Dynamic Time Warping (DTW), and Random Decision Forest (RDF), [30, 35, 4]. Meanwhile, research has been done on HMM and DTW methods to improve [11] or compare them. For example, in [29], the authors, by comparing the results obtained through the HMM and DTW methods, concluded that more training data is necessary to obtain better results from the HMM method. Also, according to this article, more data for training makes computing more complicated by this method. Therefore, the DTW has the advantage of better results with lower computational cost. In [47, 4], using the Laban Movement Analysis (LMA) [19], a robust descriptor for identifying human emotions based on their body movements is presented. This method can describe human movement using four main components: body, space, effort and shape. Because LMA is a method that takes into account all aspects of a movement, the descriptors defined in this way are highly resistant to factors such as light, background, and location. SO in this paper, we first attempt to define a robust descriptor using this method. The component effort is an element to describe how a movement is performed. Therefore, in the above articles, since the purpose is to identify emotions, it has been used. But since our goal is to recognize human movements, we have used three other components, namely, Body, Space, and Shape, to construct our descriptors. Afterwards, we try to calculate the similarities between actions using the Dynamic Time Warping algorithm. At the end, by the use of Multi Class SVM, the action are labeled. We implement proposed algorithm on four popular data-sets known as MSR Action 3D, Florence 3D actions, UTKinect-Action3D and SYSU 3D HUMAN-OBJECT INTERACTION data-sets. In the next section, we present our method and its related algorithm. In the last section, we present experiment and results compared to SoA.

2. Method

Our work is divided into two parts. In the first part, a robust descriptor is defined using the 3D coordinates of the joints. In the second part, the curves created by this descriptor, which are calculated by the use

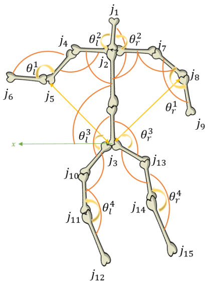


Fig. 1. Characters extracted from 3D skeleton representation for body component.

of the Dynamic Time Warping method, are trained and evaluated by the Support Vector Machine(SVM) method.

2.1. Descriptor Definition Using Laban Movement Analysis

Laban Movement Analysis (LMA) [19] is a method of describing human movements using time-dependent elements, occupied space, attitude and body orientations. The complete LMA consists of four descriptive categories, namely body, shape, space and effort, to classify the human movement. In this study, we investigate the first three categories of Body, Shape and Space:

Body component describes the structural and physical aspect of the human body and interprets the entire human body that is connected by joints and identifies which parts of the body are moving. In this article, several characters are defined for this component. The first character defined for the body element is the 3D coordinates of all the joints. Let consider $j_i = (x_i, y_i, z_i)$ is the i_{th} joint of the skeletal representation of the body, so for all joints we will have $J_n = \{j_1, j_2, \ldots, j_n\}$, where n is the number of joints and can vary depending on the type of used sensor. The next element intended for the body component is the vector between all relative joints, to do this, we compute all combinations of J_n taken 2 at a time and then the vector between them. We also added angles of body parts to the descriptor. So the angles around elbows (θ_l^1, θ_r^1), neck (θ_l^2, θ_r^2), hips (θ_l^3, θ_r^3) and knees (θ_l^4, θ_r^4) are computed by:

$$\theta^{j_j} = \arccos \frac{\overrightarrow{j_j - j_i} \cdot \overrightarrow{j_k - j_j}}{\left\| \overrightarrow{j_j - j_i} \right\| \left\| \overrightarrow{j_k - j_j} \right\|} \tag{1}$$

The ”l” index represents the left side of the skeleton and the ”r” index represents the right side of the skeleton. If we consider $j_i = (x_i, y_i, z_i)$, $j_j = (x_j, y_j, z_j)$, and $j_k = (x_k, y_k, z_k)$ as $3D$ coordinates of three consecutive joints, the angle between them will be expressed as follows: Figure 1 shows an image of the extracted characters in the Body component in a skeletal representation of 15 joints. **Space** component represents the movement of the body in relation to the its environment with spatial patterns, paths, and spatial lines. For this purpose, we calculated the curve created by a joint in two consecutive frames. So if we consider j_i^f and j_i^{f+1} two joints in the f and $f+1$ frames, we can calculate the curve created using the spline function. The inner points are a Non-linearly row vector of 50 evenly spaced points between j_i^f and j_i^{f+1}. So for all $1 \leq i \leq n$ and $1 \leq f \leq F$, where F is the total number of the frames in a motion sequence, we will have:

$$Curve(j_i) = spline(j_i^f, j_i^{f+1}) \qquad 1 \leq i \leq n \tag{2}$$

Another sub-component of the component of space that can help make this descriptor more robust is geometrical observations, whose task is to describe a movement in terms of its direction and location in its environment. For this sub-component, we calculated the quaternion of all the angles shown in red in Figure 1 in the local coordinates.
According to [7], **shape**, is a set of qualities that emerge from the Body and Space components. In general, this component itself has three sub-components namely Shape Flow, Directional Movement, and Shaping. But in this article we only use the first sub-component. Therefore, to calculate the first sub-component, for the five joints that have the highest degree of freedom of movement, respectively: head, left hand, left foot, right foot, and right hand $\{j_1, j_6, j_9, j_{12}, j_{15}\}$, we calculated the volume of polyhedron created by these joints. This step is done by calculating the volume of the convex hull of the 3D skeleton based on Quickhull algorithm [6]. This component helps us to interpret how a movement's shape changes, so we can determine whether the occupied space by the body, increases or becomes narrower.
Finally, for a data set consisting of N_v videos, in which each skeleton, according to the type of used sensor (Kinect 1 has 15 and kinect 2 has 20), has n joints, the descriptor, F, is as follows:

$$F = \left[M_{m\times D}^{1,1,1} \; M_{m\times D}^{1,1,2} \dots M_{m\times D}^{2,1,1} \dots M_{m\times D}^{a,p,r} \dots M_{m\times D}^{N_a,N_p,N_r} \right]_{N_v \times m\times D}^{T} \tag{3}$$

Where $N_v = N_a \times N_p \times N_r$ and for each $1 \leq a \leq N_a$ (**N**umber of **a**ctions), $1 \leq p \leq N_p$ (**N**umber of **p**articipants) and $1 \leq r \leq N_r$ (**N**umber of **r**epetition), we have:

$$M_{m\times D}^{a,p,r} = \left[v^1 \dots v^D \right]_{m\times D} \tag{4}$$

where for all $1 \leq d \leq D$(number of desired frames)

$$v^d = \left[v_{Position}^{1\times(3\times n)} \; v_{ArcLenght}^{1\times(n-1)} \; v_{VolumOfPolyhedron}^{1\times 1} \; v_{RelativeJoints}^{1\times(3\times C_2^{J_n})} \; v_{ThetAngles}^{1\times 8} \right]_{1\times m}^{T} \tag{5}$$

where $C_2^{J_n}$ = the number all combinations of J_n taken 2 at a time and $m = (3\times n) + (n-1) + 1 + (3\times C_2^{J_n}) + 8$.
As mentioned above, the LMA algorithm consists of four components. The fourth component, or Effort, relates to how to perform a movement that deals with the speed, acceleration, or force used to perform a gesture. Since we only deal with gestures in this article, and the way they are handled by the participants and their moods or emotions is not discussed here, we skip this component. Because if in the case of including elements such as speed, there will be a difference between a slow-moving and a fast-moving one, and the identification accuracy will decrease.

2.2. Classification Using Dynamic Time Warping

DTW is one of the common methods for measuring the similarity between two different curves. This optimization algorithm can compress or stretch the signals adaptively to create an optimal map between two time series. Using the normalized path to calculate the similarity between sequential data can overcome the problem that temporal data cannot match each other because of different signal lengths[46]. For two time series, $X = (X_1 \dots X_N)$ and $Y = (Y_1 \dots Y_M)$, their warp path can be expressed as $w_1, \dots, w_k, \dots, w_K$, $(max(N,M) \leq K \leq N+M)$. Where $w_k(a,b)$ is a link between X_a, $1 \leq a \leq N$, and Y_b, $1 \leq b \leq M$. The curves created by the algorithm must meet the following requirements:

- Boundary constraint: $w_1 = (1,1)$ and $w_K = (N,M)$.
- Monotonicity constraint: Given $w_k(a,b)$ and $w_{k+1}(a',b')$ then $a \leq a'$ and $b \leq b'$
- Continuity constraint: Given $w_k(a,b)$ and $w_{k+1}(a',b')$ then $a' \leq a+1$ and $b' \leq b+1$

Finally, the required distance according to the warped path is obtained by the following formula:

$$
\begin{aligned}
&Dist(X_a, Y_b) = (X_a - Y_b)^2 \\
&Cost_{min}(X, Y) = Dist(X_a, Y_b) + \min\{Dist(X_{a-1}, Y_b), Dist(X_a, Y_{b-1}), Dist(X_{a-1}, Y_{b-1})\}
\end{aligned}
\tag{6}
$$

Thereupon, using the DTW [23], we get the minimum cost and the warping path, namely $I_X^{1\times D}$ and $I_Y^{1\times D}$. The elements of I_Y, are the vector indicators, d, belonging to Matrix $M^{a,p,r}_{m\times D}$, that we have to go through, respectively, to achieve the minimum cost using DTW. In the next step, the matrix $M^{a,p,r}_{m\times D}$ will be updated by replacind its vector with concatenated vectors associated with the indicators in I_Y (Algorithm 1). Afterward, in order to reduce complex time series data, as it is proposed in [25], a Fast Fourier Transform (FFT) is applied to all curves obtained by the DTW. According to the [16], SMV is a powerful classifier for classifying FFT-based data. In general, SVM is a binary classifier, but it can also be used as a multi-class classifier. LIBSVM [12], which is a most widely used tool for SVM classification is used to label training data. The following figure 2 is an overview of all the steps involved:

Algorithm 1 Used algorithm.

```
 1. Input: $F$
 2. Output: Action labels
 3. # Divide $F$ into two parts, training, $F_{train}$ and validation, $F_{test}$.
 4. For $1 \le a \le N_a$
 5.     For $1 \le p \le N_p$
 6.         If $M^{a,p,r}_{m\times D} \in F_{train}$
 7.             $F^a_{train} \leftarrow M^{a,p,r}_{m\times D}$
 8.         Else If
 9.             $F^a_{test} \leftarrow M^{a,p,r}_{m\times D}$
10.         End If
11.     End
12.     $ReferenceCurve \leftarrow$ The first $M^{a,p,r}_{m\times D}$ in $F^a_{train}$
13.     $Repetition \leftarrow 0$
14.     While ($Repetition \le$Threshold)          ▶ The threshold varies according to the data sets.
15.         For $1 \le h \le s^a_{train}$          ▶ $s^a_{train}$ = the total number of $M^{a,p,r}_{m\times D}$ in $F^a_{train}$.
16.             $Cost_{min}, I_X, I_Y \leftarrow$DTW$_{element-wise}(ReferenceCurve^T, F^a_{train}(h)^T)$
17.             For $1 \le d \le D$
18.                 $Curve(h) \leftarrow$ Replace the $d_{th}$ column of $F^a_{train}(h)$ with $v^{I_Y(d)}$
19.             End
20.         End
21.         $RefrerenceCurve \leftarrow$ standard deviation of the elements of $Curve$ along $h$.
22.         $Repetition \leftarrow Repetition + 1$
23.     End
24.     $Curve \leftarrow$ FastFourierTransforms($Curve$)
25.     $M_{Train} \leftarrow Curve$
26. End
27. Train the $M_{Train}$ using Support Vector Machine.
28. Repeat steps 12 to 26 for $F^a_{test}$ to obtain the validation matrix, $M_{test}$.
29. Label samples in $M_{test}$ using Support Vector Machine.
```

3. Experiment and Results

In this section, we evaluate the proposed descriptor that we defined using the LMA method on four public data-sets using Dynamic Time Warping algorithm.

As mentioned above, the proposed method in this paper has been evaluated on four general data-sets

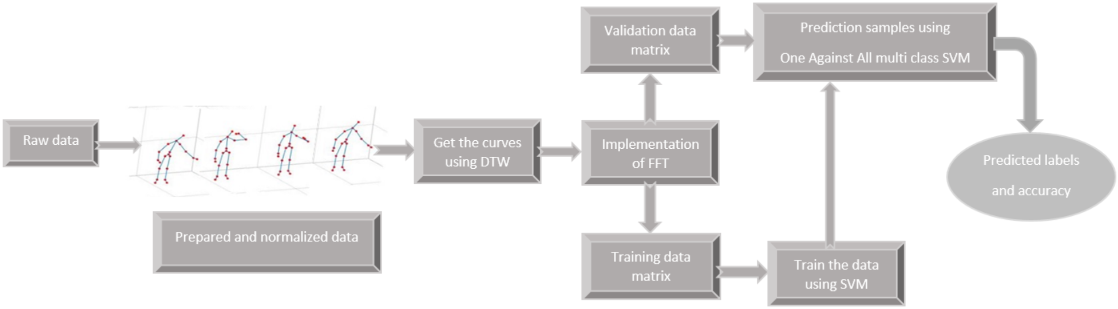


Fig. 2. An overview of all the steps involved.

namely MSR Action 3D, UTKinect-Action3D, Florence 3D actions and SYSU 3D HUMAN-OBJECT INTERACTION data-sets. The software used to produce the experimental results is MATLAB 2018b.
Before showing the results, a brief description of each data set is given below.

*3.1. **Used Data-sets***

MSR Action 3D data-set is a public data-set [21] that many researchers have evaluated their proposed methods by implementing on it. This data-set includes 20 actions performed by 10 subjects facing a Kinect device. Each action is performed 2 or 3 times by the participants. In total, 567 sequences are available. For each sequence, the data-set provides depth information, color information and skeleton information. In our case, we only use the skeleton data. As reported in [36], 10 actions are not used in the experiments because the skeletons are either missing or too erroneous. For our experiments, we use 557 sequences.

UTKinect Action data-set [42] is another challenging public data-set, including 10 different action performed by 10 different subjects. Each action is performed 2 or 3 times by the participants, In total, 200 sequences are available.

Florence-3D data-set [5], this is also a public data-set that contains 9 different actions. Each action is repeated two or three times by 10 different participants.In total, 215 sequences are available.

SYSU 3D HUMAN-OBJECT INTERACTION data-set[14] in this data set, 40 participants were asked to do 12 daily activities. In each of these gestures, participants interacted with six different objects: phone, chair, bag, wallet, mop and besom. So there are a total of 480 sequences in this data set, each sequence using the Kinect sensor gives the information about the RGB frames, depth sequence and 3D coordinates of skeleton data.

3.2. Development and Result

3.2.1. Normalization

The first step is to normalize raw data. Normalization of data is important because an action in a data set is performed by different people of different sizes and also in different primary locations. So in the case of non-normalization, the proposed descriptor which is based on the three-dimensional coordinates of the joints and also vectors between successive joints and their size have been defined, cannot be invariant, and this can have an effect on reducing their robustness, which reduces gesture recognition accuracy. to make the skeletal data invariant to the body size of the subjects, each component of all vectors between two consecutive joints is divided by its magnitude. Also, in order to make the skeleton data invariant to the angle of each person relative to the used sensor (which is Microsoft Kinect Sensor in all three data-sets used in this work). We matched the X axis of hip center to the X axis of Kinect axis using the rotation matrix.

3.2.2. Data Sampling

The next step in preparing the data for use in the algorithm described in the previous section is to fix the number of frames per gesture. For this purpose, the data should be sampled according to the number of desired frames. To do this, the spline function is used. In figure 3, the path traveled by the head joint is shown in the different movements performed by different persons. The red path represents the path with

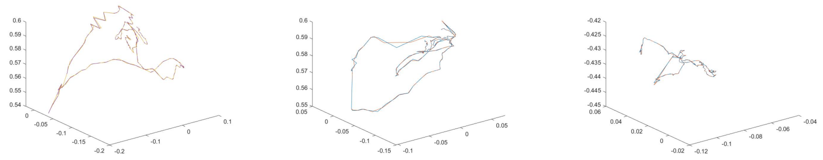


Fig. 3. Red: the path traveled by the head joint in different action performed by different subject with the real number of frames and blue with the desired frames.

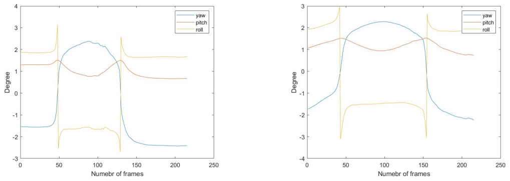


Fig. 4. (a) Yaw, Pitch and Roll Changes for Person30, Action7; (b) Yaw, Pitch and Roll Changes for Person40, Action7.

the real number of frames and the blue path represents the path with the number of desired frames. As can be seen using the spline interpolation function, the paths are perfectly consistent.

3.2.3. *Angle Representation*

In the last step, in order to evaluate the efficiency of the proposed quaternion angles, we calculated the triple-angle of Roll, Pitch and Yaw. As shown in figure 4, an action preformed by two different participants has the same amplitude variations.

3.3. *Comparison With Sate of Art and Discussion*

According to the state of art, there are two ways to use the MSR-Action 3D data-set, in the first way, the data-set is used as a whole. we used 60% of the data for training and 40% for testing, the data are randomly divided. Using this method, the calculated average accuracy is 90.55%, as shown in the following table 1, this value is superior to similar work in the field of action recognition based on skeletal data. Due to the existence of large amount of computation in training and validation phases that may lead to decrease the accuracy, the author in [21], have proposed to divide the whole data-set into three different action categories (AS1,AS2,AS3), each consisting of 8 actions.

Table 1. Comparison with the state-of-the-art results MSR-Action3D data-set.

Method	Year	Accuracy(%)
Active Joints [34]	2107	84.72
EigenJoints[45]	2012	82.30
AHON4D [27]	2013	88.36
Cooperative Warp [32]	2019	90.90
Actionlet Ensemble [37]	2012	88.20
Coding Kendall's Shape Trajectories [8]	2018	86.18
Learning Composite Latent Structures[41]	2019	87.2
HAR using CNN[2]	2019	87.1
Our method	2020	90.55

By implementing the second method, the recognition accuracy has reached 99.24 (Average accuracy of AS1, AS2 and AS3). Given that the size of each category is small, we used only 30% of the data for validation, and 70% of it was allocated to the training part. In AS1 and AS2, most movements are related to the upper and middle trunk, and in some cases the movements are very similar, for example drawing x and

Table 2. Comparison with the state-of-the-art results AS1/AS2/AS3.

Method	Year	AS1 Accuracy(%)	AS2 Accuracy(%)	AS3 Accuracy(%)
Mining Key Skeleton Poses with Latent SVM[22]	2017	89.1	88.7	94.9
Lie Group using deep network [28]	2018	96.64	87.52	98.71
LMA Qualities [3]	2019	90.3	88.7	93.1
Improving bag-of-poses[1]	2019	94.3	94.6	97.7
Coding Kendall's Shape Trajectories [8]	2018	95.87	86.72	100
DMM-UDTCWT [32]	2019	95.6	93.82	96.6
Our method	2020	99.13	98.60	100

Table 3. Comparison with the state-of-the-art results UTKinect Action, Florence 3D Accurancy and SYSU 3D HOI.

Method	Year	UTKinect Action Acc(%)	Florence 3D Acc(%)	SYSU 3D HOI Acc(%)
Active Joint [34]	2017	95.96	-	-
Mining Key Skeleton Poses with Latent SVM[22]	2017	91.5	87	-
Motion Trajectories [13]	2012	91.50	87.0	-
Cooperative Warp [32]	2019	95.38	88.38	-
Grassmann manifold [31]	2015	88.50	-	-
Group Sparse Regression[20]	2018	95.1	-	80.7
Reinforcement Learning[33]	2018	-	-	76.9
Self-Attention Guided Deep Features[43]	2019	-	-	80.36
HRS networks[44]	2019	-	-	84.23
Traj. on $S^+(3,n)$- BP Fusion[17]	2018	96.48	-	80.22
SVRNN[9]	2019	89.0	-	-
Physiological function assessment [11]	2019	-	-	83.75
Geometric Algebra Representation [10]	2019	-	-	84.62
Progressive Teacher-student Learning [39]	2019	-	-	87.92
Deep Bilinear Learning [15]	2018	-	-	88.9
Our method	2020	97.36	94.22	Setting 1:86.63/2: **92.32**

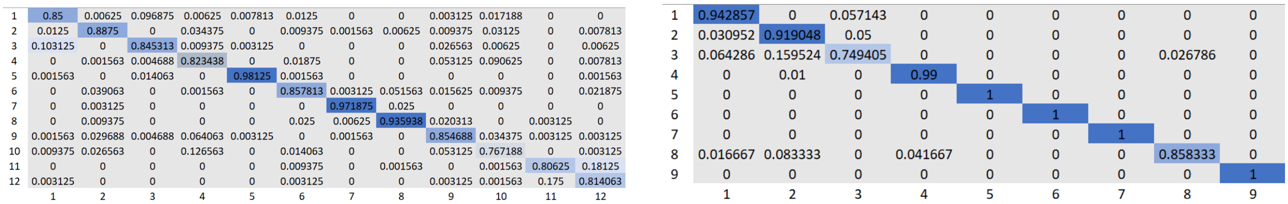
ticking. Therefore, it is reasonable to have a low accuracy compared to AS3, which consists of gestures in which all body organs are involved and are also very different. The results of this method and comparison with previous work can be seen in the following table 2. The results obtained from the UTKinect Action, 60% training and 40% testing, Florence 3D and SYSU 3D HOI data-sets, 60% training and 40% testing, respectively, are as follows (Table 3): As we can see, our proposed method outperforms most of the work in the state of art.

Since the size of SYSU 3D HOI is almost the same as the MSR-Action3D, it can be concluded that in this data-set, high calculations in the classification may reduce the accuracy. So we decided to split this data-set into two categories. Drinking, Calling phone, Wearing backpacks, Sitting chair, Taking out wallet and Mopping are placed in the first category (**AC1**), and Pouring, Playing phone, Packing backpacks, Moving chair, Taking from wallet and Sweeping are in the second category (**AC2**). After implementing this division, allocating 40% of the data for validation and 60% for training, the accuracy increased to **92.32%** (average accuracy of **AC1 (92.31%)** and **AC2 (92.33%)**).

And in the figure 5, the confusion matrix for the SYSU HOI and Florence data-sets are shown.

4. Conclusion and Future Work

In this paper, we have introduced a robust and invariable descriptor using the Laban Movement Analysis method, since our focus was only on gesture recognition, so only three of the four components of this method were used. In this approach, elements must be selected that reflect the changes in the body as it moves. For example, one of the elements added to this article is the angle around the hip. Since the data-sets tested here often include gestures that require all parts of the body to perform, this angle, is a



1	0.85	0.00625	0.096875	0.00625	0.007813	0.0125	0	0	0.003125	0.017188	0	0
2	0.0125	0.8875	0	0.034375	0	0.009375	0.001563	0.00625	0.009375	0.03125	0	0.007813
3	0.103125	0	0.845313	0.009375	0.003125	0	0	0	0.026563	0.00625	0	0.00625
4	0	0.001563	0.004688	0.823438	0	0.01875	0	0	0.053125	0.090625	0	0.007813
5	0.001563	0	0.014063	0	0.98125	0.001563	0	0	0	0	0	0.001563
6	0	0.039063	0	0.001563	0	0.857813	0.003125	0.051563	0.015625	0.009375	0	0.021875
7	0	0.003125	0	0	0	0	0.971875	0.025	0	0	0	0
8	0	0.009375	0	0	0	0.025	0.00625	0.935938	0.020313	0	0.003125	0
9	0.001563	0.029688	0.004688	0.064063	0.003125	0	0.001563	0	0.854688	0.034375	0.003125	0.003125
10	0.009375	0.026563	0	0.126563	0	0.014063	0	0	0.053125	0.767188	0	0.003125
11	0	0	0	0	0	0.009375	0	0.001563	0	0.001563	0.80625	0.18125
12	0.003125	0	0	0	0	0.003125	0	0	0.003125	0.001563	0.175	0.814063
	1	2	3	4	5	6	7	8	9	10	11	12

1	0.942857	0	0.057143	0	0	0	0	0	0
2	0.030952	0.919048	0.05	0	0	0	0	0	0
3	0.064286	0.159524	0.749405	0	0	0	0	0.026786	0
4	0	0.01	0	0.99	0	0	0	0	0
5	0	0	0	0	1	0	0	0	0
6	0	0	0	0	0	1	0	0	0
7	0	0	0	0	0	0	1	0	0
8	0.016667	0.083333	0	0.041667	0	0	0	0.858333	0
9	0	0	0	0	0	0	0	0	1
	1	2	3	4	5	6	7	8	9

Fig. 5. Average confusion matrix of (a) SYSU HOI and (b)Florence 3D-Action data-set.

good connection for connecting the upper and the lower trunk. Also, in this article, it was shown that by dividing the SYSU 3D HOI data-set into two part (AC1 and AC2), accuracy can be increased for 5.69%. In the second part, the performance of the proposed descriptor by the use of dynamic time warping method was tested. Four public data-sets were investigated. The obtained results in our method were in many cases superior to other studies in gestures recognition based on 3D skeletal coordinates. The lowest accuracy rate was related to the MSR Action 3D data-set, which, after dividing it into three groups, revealed that there were many similar movements in one group ($AS2$), and that the same movements in the overall data set caused confusion. By adding factors such as speed, acceleration and force to the descriptor presented in this article, it can also be used to identify emotions.
In the next work, we try to use this descriptor to identify emotions based on body movements.

References

[1] Agahian, S., Negin, F., Köse, C., 2019. Improving bag-of-poses with semi-temporal pose descriptors for skeleton-based action recognition. The Visual Computer 35, 591–607.

[2] Ahmad, Z., Illanko, K., Khan, N., Androutsos, D., 2019. Human action recognition using convolutional neural network and depth sensor data, in: Proceedings of the 2019 International Conference on Information Technology and Computer Communications, pp. 1–5.

[3] Ajili, I., Mallem, M., Didier, J.Y., 2019a. Human motions and emotions recognition inspired by lma qualities. The Visual Computer 35, 1411–1426.

[4] Ajili, I., Ramezanpanah, Z., Mallem, M., Didier, J.Y., 2019b. Expressive motions recognition and analysis with learning and statistical methods. Multimedia Tools and Applications , 1–26.

[5] Bagdanov, A.D., Del Bimbo, A., Masi, I., 2011. The florence 2d/3d hybrid face dataset, in: Proceedings of the 2011 joint ACM workshop on Human gesture and behavior understanding, pp. 79–80.

[6] Barber, C.B., Dobkin, D.P., Dobkin, D.P., Huhdanpaa, H., 1996. The quickhull algorithm for convex hulls. ACM Transactions on Mathematical Software (TOMS) 22, 469–483.

[7] Bartenieff, I., Lewis, D., 1980. Body movement: Coping with the environment. Psychology Press.

[8] Ben Tanfous, A., Drira, H., Ben Amor, B., 2018. Coding kendall's shape trajectories for 3d action recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2840–2849.

[9] Butepage, J., Kjellstrom, H., Kragic, D., 2019. Predicting the what and how-a probabilistic semi-supervised approach to multi-task human activity modeling, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 0–0.

[10] Cao, W., Lu, Y., He, Z., 2019a. Geometric algebra representation and ensemble action classification method for 3d skeleton orientation data. IEEE Access 7, 132049–132056.

[11] Cao, W., Zhong, J., Cao, G., He, Z., 2019b. Physiological function assessment based on kinect v2. IEEE Access 7, 105638–105651.

[12] Chang, C.C., 2011. " libsvm: a library for support vector machines," acm transactions on intelligent systems and technology, 2: 27: 1–27: 27, 2011. http://www. csie. ntu. edu. tw/˜ cjlin/libsvm 2.

[13] Devanne, M., Wannous, H., Berretti, S., Pala, P., Daoudi, M., Del Bimbo, A., 2014. 3-d human action recognition by shape analysis of motion trajectories on riemannian manifold. IEEE transactions on cybernetics 45, 1340–1352.

[14] Hu, J.F., Zheng, W.S., Lai, J., Zhang, J., 2015. Jointly learning heterogeneous features for rgb-d activity recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 5344–5352.

[15] Hu, J.F., Zheng, W.S., Pan, J., Lai, J., Zhang, J., 2018. Deep bilinear learning for rgb-d action recognition, in: Proceedings of the European Conference on Computer Vision (ECCV), pp. 335–351.

[16] Islam, S.M., Rahman, A., Prasad, N., Boric-Lubecke, O., Lubecke, V.M., 2019. Identity authentication system using a support vector machine (svm) on radar respiration measurements, in: 2019 93rd ARFTG Microwave Measurement Conference (ARFTG), IEEE. pp. 1–5.

[17] Kacem, A., 2018. Novel Geometric Tools for Human Behavior Understanding. Ph.D. thesis.

[18] Kołakowska, A., 2013. A review of emotion recognition methods based on keystroke dynamics and mouse movements, in: 2013 6th International Conference on Human System Interactions (HSI), IEEE. pp. 548–555.

[19] von Laban, R., 1950. The Mastery of Movement on the Stage.(1. Publ.). Macdonald and Evans.

[20] Li, M., Yan, L., Wang, Q., 2018. Group sparse regression-based learning model for real-time depth-based human action prediction. Mathematical Problems in Engineering 2018.

[21] Li, W., Zhang, Z., Liu, Z., 2010. Action recognition based on a bag of 3d points, in: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops, IEEE. pp. 9–14.

[22] Li, X., Zhang, Y., Liao, D., 2017. Mining key skeleton poses with latent svm for action recognition. Applied Computational Intelligence and Soft Computing 2017.

[23] MATLAB, 2018. 9.7.0.1190202 (R2019b). The MathWorks Inc., Natick, Massachusetts.

[24] Men, Q., Leung, H., Yang, Y., 2019. Self-feeding frequency estimation and eating action recognition from skeletal representation using kinect. World Wide Web 22, 1343–1358.

[25] MOHAMMED, M.Y.Y., CELIK, M., 2019. Developing fast techniques for periodicity analysis of time series, in: 2019 3rd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), IEEE. pp. 1–5.

[26] Nunez, J.C., Cabido, R., Pantrigo, J.J., Montemayor, A.S., Velez, J.F., 2018. Convolutional neural networks and long short-term memory for skeleton-based human activity and hand gesture recognition. Pattern Recognition 76, 80–94.

[27] Oreifej, O., Liu, Z., 2013. Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 716–723.

[28] Rhif, M., Wannous, H., Farah, I.R., 2018. Action recognition from 3d skeleton sequences using deep networks on lie group features, in: 2018 24th International Conference on Pattern Recognition (ICPR), IEEE. pp. 3427–3432.

[29] Sajjan, S.C., Vijaya, C., 2012. Comparison of dtw and hmm for isolated word recognition, in: International Conference on Pattern Recognition, Informatics and Medical Engineering (PRIME-2012), IEEE. pp. 466–470.

[30] Sinha, K., Kumari, R., Priya, A., Paul, P., 2019. A computer vision-based gesture recognition using hidden markov model, in: Innovations in Soft Computing and Information Technology. Springer, pp. 55–67.

[31] Slama, R., Wannous, H., Daoudi, M., Srivastava, A., 2015. Accurate 3d action recognition using learning on the grassmann manifold. Pattern Recognition 48, 556–567.

[32] Sun, Z., Guo, X., Li, W., Liu, Z., 2019. Cooperative warp of two discriminative features for skeleton based action recognition, in: Journal of Physics: Conference Series, IOP Publishing. p. 042027.

[33] Tang, Y., Tian, Y., Lu, J., Li, P., Zhou, J., 2018. Deep progressive reinforcement learning for skeleton-based action recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5323–5332.

[34] Tehrani, A.K., Aghbolaghi, M.A., Kasaei, S., 2017. Skeleton-based human action recognition .

[35] Vemulapalli, R., Arrate, F., Chellappa, R., 2014. Human action recognition by representing 3d skeletons as points in a lie group, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 588–595.

[36] Wang, C., Wang, Y., Yuille, A.L., 2016. Mining 3d key-pose-motifs for action recognition, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

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

[38] Wang, P., Li, W., Li, C., Hou, Y., 2018. Action recognition based on joint trajectory maps with convolutional neural networks. Knowledge-Based Systems 158, 43–53.

[39] Wang, X., Hu, J.F., Lai, J.H., Zhang, J., Zheng, W.S., 2019. Progressive teacher-student learning for early action prediction, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3556–3565.

[40] Wei, H., Jafari, R., Kehtarnavaz, N., 2019a. Fusion of video and inertial sensing for deep learning–based human action recognition. Sensors 19, 3680.

[41] Wei, P., Sun, H., Zheng, N., 2019b. Learning composite latent structures for 3d human action representation and recognition. IEEE Transactions on Multimedia 21, 2195–2208.

[42] Xia, L., Chen, C.C., Aggarwal, J.K., 2012. View invariant human action recognition using histograms of 3d joints, in: 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, IEEE. pp. 20–27.

[43] Xiao, R., Hou, Y., Guo, Z., Li, C., Wang, P., Li, W., 2019. Self-attention guided deep features for action recognition, in: 2019 IEEE International Conference on Multimedia and Expo (ICME), IEEE. pp. 1060–1065.

[44] Xie, C., Li, C., Zhang, B., Pan, L., Ye, Q., Chen, W., 2019. Hierarchical residual stochastic networks for time series recognition. Information Sciences 471, 52–63.

[45] Yang, X., Tian, Y.L., 2012. Eigenjoints-based action recognition using naive-bayes-nearest-neighbor, in: 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, IEEE. pp. 14–19.

[46] Yu, L., Xiong, D., Guo, L., Wang, J., 2016. A remote quantitative fugl-meyer assessment framework for stroke patients based on wearable sensor networks. Computer methods and programs in biomedicine 128, 100–110.

[47] Zacharatos, H., Gatzoulis, C., Chrysanthou, Y., Aristidou, A., 2013. Emotion recognition for exergames using laban movement analysis, in: Proceedings of Motion on Games, ACM. pp. 61–66.