Abstract—Motion recognition from videos is actually a very complex task due to the high variability of motions. This paper describes the challenges of human motion recognition, especially motion representation step with relevant features. Our descriptor vector is inspired from Laban Movement Analysis method. We propose discriminative features using the Random Forest algorithm in order to remove redundant features and make learning algorithms operate faster and more effectively. We validate our method on MSRC-12 and UTKinect datasets.

Keywords—Human motion recognition, Discriminative LMA features, random forest, features reduction.

I. INTRODUCTION

HUMAN motion recognition is a fundamental topic in the field of computer vision. One of the main challenges with motion recognition is that the same motion may be performed in different ways by different persons, and even by the same person. Although a significant amount of research has been focused on human motion representation, it remains still not enough. First motion features proposed in the computer vision literature were based on interest points [8], [11]. After, another type of features were studied based on depth informations provided from depth sensors. The depth cameras in general provide better quality 3D data than those estimated from monocular video sensors. This allows to focus on the analysis of the human motion through the identification of his joints [7], [16], [12], [10]. However, such proposed features did not take into account the semantic aspects of motion, for example to classify two actions with the same movement but did not take into account the semantic aspect of motion. Another key challenge in motion representation is to realize a good compromise between robustness performance and computational costs. The assumption that increasing the number of features can provide more informations about motion is not always valid in practice, because it can be time consuming and may lead to finding a less optimal solution. For feature reduction, several approaches were proposed and could be divided in two groups: methods based on statistical measures [5], [9], and methods based on learning algorithms [13], [15]. The first category consists in ranking features according to some statistical measures. It is fast and independent of any classifier, but it requires a threshold to select the top ranked features. Finally, in statistical approaches, some important features that are less informative on their own, but they are informative when combined with others can be discarded. The second category evaluates the importance of a random subset of features by training a model on it. A learning method is used to evaluate the importance of each combination of features on the classification performance. Despite the effectiveness of these methods, they have the constraint to be computationally more expensive compared to the statistics methods due to the repeated learning and cross validation steps.

In this paper, which extends our preliminary work presented in [1], [2], we address the problem of analyzing human motion from skeleton sequences captured by depth cameras. Particularly, our work focuses on representing human motions by keeping most salient and complementary features based on Random Forest algorithm. Our descriptor vector is inspired from Laban Movement Analysis method (LMA) to describe quantitative and qualitative representations of motions. The rest of the paper is structured as follows. Section 2 describes our proposed approach with motion recognition steps. Section 3 presents the experimental results on MSRC-12 and UTKinect datasets. Finally, conclusions and future work are stated in Section 4.

II. PROPOSED APPROACH

A. Data Acquisition

We use kinect sensor for data acquisition to extract 3D skeleton joints in real time. The first step is the normalization of all skeletons which consists in aligning all skeletons in the center of the kinet coordinate system with the base $B$ at initial frame. Given a motion sequence $S = \{J_{t}\}$, $j \in 1, \ldots, N, t \in 1, \ldots, T$, $J_{jt}$ corresponds to the coordinates of the joint $j$ captured at frame $t$. We define a local coordinate system to the skeleton anchored to the hip center joint ($J_h$), represented by the base $B'$, equipped with three unit vectors, the left hip joint vector $\vec{n}_{lh}$, the spine vector $\vec{n}_s$ and their cross product $\vec{n}_c = \vec{n}_{lh} \wedge \vec{n}_s$. For each sequence, we first apply translation to move the skeleton to the center of kinet, and after a rotation to align both coordinate systems (Fig. 1). The transformed joint yields to:

\[ J_{j,t}B' = R_{B \rightarrow B'}^{-1}(J_{j,t}B - J_{c,t}B) \]  (1)

\[ R_{B \rightarrow B'} = \left[ \begin{array}{c} \vec{n}_{lh}^T \\ \vec{n}_s^T \\ \vec{n}_c^T \end{array} \right] \]  (2)

\[ \vec{n}_{lh} = [J_{h,t+1}B - J_{e,t-1}B] \]  (3)

\[ \vec{n}_s = [J_{s,t+1}B - J_{c,t+1}B] \]  (4)
Once we applied transformations to all sequences, our system is independent of the initial position and orientation of the subject in the scene. Then, we pass to the next step which consists in converting the skeleton joints data to a descriptor vector based on LMA method.

![Kinect coordinate system]

**B. Motions Representation**

LMA approach employs a multilayered description of movement, focusing on four components: Body, Space, Shape, and Effort. **Body** component has the responsibility of highlighting the body part which is moving, making the connection between the moving parts and taking in consideration the issues of locomotion and kinematics. For this category, we describe the organization and connection between the different joints (Fig. 2). We consider two parts, the upper and lower part. For the first one, the extension of the different joints is described by computing the following angles in left and right parts respectively: between hands and shoulders ($\theta_1$, $\theta_1'$), between elbows and hips ($\theta_2$, $\theta_2'$), between elbows and shoulders in the symmetrical part ($\theta_3$, $\theta_3'$). We also calculate the distances between the two hands ($d_{HH}$) as well as the distances between the shoulder center and both hands ($d_{shc lh}$, $d_{shc rh}$). It allows us to have a more idea on the way of the two hands. For the lower part of the body, the extension of the knees has been described with the angles between the feet and the hips ($\theta_4$, $\theta_4'$). These two characteristics allow to characterize specific actions like crouch or hide gestures. We also characterize the opening of the legs with the angle computed between the two knees ($\theta_{ls}$). Mean, standard deviation, and range of the body features are computed to quantify the Body component. We compute the length of trajectories ($L$) made by upper and lower body extremities (head, hands, and feet) to quantify the **Space** component.

$$L = \sum_{t=1}^{T-1} ||J_{j,t+1} - J_{j,t}||$$

In **Shape** component, we describe the way the body changes shape during movement with three qualities. In the first **Shape flow** factor, we characterize the change shape in a self-to-self relationship by computing the volume of the smallest convex envelope of the human body based on Quickhull algorithm [3], as shown in Fig. 2. The second factor is the **Directional movement**, we define the pathway of the movement of upper body extremities (hands and head) through space by computing their curvatures ($C$).

$$C = \sum_{t=2}^{T-1} \arccos\left(\frac{\langle J_{j,t-1}, J_{j,t} \rangle}{\|J_{j,t-1}\| \|J_{j,t}\|}\right)$$

Finally, we quantify the **Carving** factor of the **Shape** component which describes the qualitative changes in the shape relating to spine joint pose at initial frame $J_{s,1}$, according to three planes: Horizontal ($D_H$), Frontal ($D_F$), and Sagittal ($D_S$), relating them to bipolar descriptors: spreading/enclosing, rising/sinking, and retreating/advancing, respectively (Fig. 2).

$$D_H = \frac{1}{T} \sum_{t=1}^{T} \sum_{j=1}^{N} \sqrt{(J_{x,j,t} - J_{x,s,1})^2}$$

$$D_F = \frac{1}{T} \sum_{t=1}^{T} \sum_{j=1}^{N} \sqrt{(J_{y,j,t} - J_{y,s,1})^2}$$

$$D_S = \frac{1}{T} \sum_{t=1}^{T} \sum_{j=1}^{N} \sqrt{(J_{z,j,t} - J_{z,s,1})^2}$$

**Effort** component describes how the body concentrates its effort while performing a motion and characterizes expressive behaviors based on four factors: **Time**: Sudden/Sustained, **Weight**: Light/Strong, **Flow**: Bound/Free, and **Space**: Direct/Indirect (Fig. 3). In Effort component we focus on the upper body part (head, hands, and spine), since it was the most expressive part during human motion. Joints velocities and accelerations are computed for quantifying **Time** and **Weight** factors, respectively. Three measures of variability (Mean, standard deviation, and range) are used for both features. We quantify the **Flow** factor by computing the yaw and pitch range of joints motion. For Free motion we will obtain a higher range compared to Bound motion. To describe the direction of the movement in space for **Space** factor, we compute the Straightness index ($S$) of joints motion as the ratio of the distance between the first and last frame ($D$) to the sum of the displacements between two successives frames ($L$).

$$S = \frac{D}{L}$$

C. **Motions Recognition**

For motions training and classification, we apply the Random Forest approach (RF) [4]. This method consists of an ensemble of decision trees, each tree is grown using by a different bootstrap sample from the training data. Let the feature vector be $v = \{f_i\}$, $i = 1, \ldots, d$, where $d$ is the number of features for each sample. At each node, best split is chosen from a random sample of $p$ features from $d$. Consider a node $k$ comprising $S_k$ samples, split into left and right child nodes with subsamples of $S_{kl}$ and $S_{kr}$, respectively, the tree is then grown by selecting the splitting condition that maximizes the purity of the resulting tree. Gini index $I(S_k)$ is
used to select the feature at each internal node $k$. The amount of homogeneity gain achieved by the splitting node $k$ in feature $f$ can be evaluated in the following equation:

$$G(f, S_k) = I(S_k) - \sum_{i \in \{l, r\}} \left( \frac{|S_{ki}|}{|S_k|} I(S_k) \right)$$

(11)

where $I(S_k) = 1 - \sum_{j=1}^{l} \left( \frac{S^j_k}{N_k} \right)^2$, $l$ is the number of classes in node $k$, $S^j_k$ denotes the number of learning samples which belong to class $j$ at node $k$. Therefore, after several selections for $f$, the one producing the lowest value of Gini index is picked as the split criterion for the node. In the testing step, each test sample is simultaneously pushed through all trees, starting from the root, and assigning the data to the right or left child recursively until a leaf node is reached. Finally, the forest chooses the classification having the majority of votes from each of the decision trees made.

**D. Features Reduction**

Feature reduction step consists in keeping the smallest subset of most relevant features for motions representation to achieve a good compromise between accuracy and runtime in the classification process. The measure of relevant feature is achieved using RF method. During training phase, each tree is grown using a different bootstrap sample from the original training data, leaving 1/3 as OOB (Out Of Bag) to estimate the prediction error of OOB. The importance of the feature $f_t$ is measured as the difference between OOB error rate of each tree before and after permuting $f_t$.

$$I^t(f_t) = \sum_{j \in O^t} \frac{I(y_j \neq y_{j,t})}{|O^t|} - \sum_{j \in O^t} \frac{I(y_j \neq y_{j,t}^*)}{|O^t|}$$

(12)

$$I(f_t) = \frac{1}{T} \sum_{t=1}^{T} I^t(f_t)$$

(13)

$O^t$ corresponds to OOB samples for a tree $t$. $y_{j,t}$ and $y_{j,t}^*$ are the predicted classes for the $j^{th}$ sample before and after permuting the feature $f_t$ respectively. Finally, a high decrease in accuracy is an indication of the feature importance. We start with the whole set of features, we compute and record the OOB error rate. After we sort the features in descending order of importance, and we remove the feature of small importance $f_{min}$. Moreover, we apply the Tukey’s test ($\alpha = 0.05$) to simultaneously remove features that do not give a significant difference of the OOB error rate result.

**Algorithm 1: Feature reduction process**

```
Input : $v_0 = \{f_1, \ldots, p\}$ > $v_0$ is the whole feature set.
Output: $v' = \{f_j\}$, $j = 1, \ldots, p'$ > $v'$ subset of most relevant features.
1. $k = 0$
2. while $|v_k| \geq 1$
3. for $i = 1$ to $p$
4. Compute $I(f_i)$ > Importance of each feature in $v_k$.
5. end
6. Sort $\{f_i\}$ in descending order according to values of $I(f_i)$
7. $f_{min} = \text{argmin}\{I(f_i)\}$
8. Apply Tukey’s test and select set of features $\{f_i\}$ that does not lead to a significant changement of $E_k$
9. $R = f_{min} \cup \{f_i\}$
10. $v_{k+1} = v_k \setminus R$
11. $k = k + 1$
12. end
13. $v' = \text{argmin}\{E_k(v_k)\}$ > $v'$ is the optimal feature subset with minimal OOB error.
```

**III. EXPERIMENTAL RESULTS**

To evaluate the performance of our method, we use two public action datasets, MSRC-12 [6] and UTKinect [14], we report two measures: the mean of fscores and the OOB error rate. For the first measure, we adopt the 5-fold cross
validation to optimize the RF parameters, and compute the averaged results. We employ the commonly used F-score as the performance measure.

1) Evaluation on MSRC-12 Dataset: MSRC-12: is a dataset composed of 594 sequences, containing the performances of 12 gestures by 30 people. In total, there are 6244 gesture instances. The gesture classes are divided into two groups: metaphoric gestures, and iconic gestures. The motion files contain 3D coordinates of 20 joints captured at a sample rate of 30Hz.

We converted the raw data into a descriptor vectors based on our LMA qualities. Our descriptor vector composed of 85 features was fed into a learning algorithm, RF. Most important parameters of RF, the number of trees \( n_{trees} \) and the number of features to consider when splitting a node \( \text{max features} \) were adjusted. We varied \( n_{trees} \) starting from 10 until 200 trees, and we tested three values of \( \text{max features} = 85, \log_{2}(85), \) and \( \sqrt{85} \). Best recognition rate of 94.89% was achieved when setting \( n_{trees} = 100 \) and \( \text{max features} = \log_{2}(85) \), and almost the same recognition result of 94.88% was obtained for \( n_{trees} = 100 \) and \( \text{max features} = \sqrt{85} \). We also confirmed the RF parameters values by computing OOB error rate while varying \( n_{trees} \) and \( \text{max features} \). As we can see in Fig. 4, the two curves of \( \text{max features} = \log_{2}(85) \), and \( \sqrt{85} \) are very close with a very little superiority result of \( \text{max features} = \log_{2}(85) \). We identified the minimum value of \( n_{trees} \) where OOB error stabilize (around 0.008), we found \( n_{trees} = 100 \), which confirms the result obtained with the recognition rate measure. Table I illustrates the recognition results of our method compared to the state of the art methods on MSRC-12 dataset.

Our method outperforms the state of the art methods, and is very close to the result obtained in [12], which confirms the robustness of our descriptor in characterizing both iconic and metaphoric gestures with an accuracy rates of 99% and 93%, respectively. After evaluating our descriptor vector on gestures recognition in MRC12-dataset, we applied our feature reduction algorithm (Algorithm 1) in order to keep only most discriminant features according to this dataset. To obtain more stable results and better estimations for the expected OOB error rate, we repeated this procedure 30 times and average the results. In Table II, we make a comparison between results obtained before and after feature reduction process, in terms of Mean F-score, OOB error rate \( err_{OOB} \), and number of features \( N \). We notice that the number of relevant features is decreased about 20% achieving the same F-score results and decreasing the OOB error rate. We obtained a low OOB error rate of 0.004 with a number of relevant features \( N = 67 < 85 \) (see Fig. 5).

2) Evaluation on UTKinect Dataset: We also evaluated our descriptor with UTKinect dataset which is composed of 10 subjects performing 10 different activities in varied views namely walk, sit down, stand up, pick up, carry, throw, push, pull, wave hands, and clap hands. A total number of 199 sequences are available. Each action is repeated twice by the actor. Sequences are captured using one Kinect in indoor settings and their length ranges from 5 to 120 frames. This is a challenging dataset due to variations in the view point and high intra-class variations where each actor performs actions in different views. We applied both methods, feature

<table>
<thead>
<tr>
<th>Methods</th>
<th>Recognition rates (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hussein et al. [7]</td>
<td>91.70</td>
</tr>
<tr>
<td>Zhou et al. [16]</td>
<td>90.22</td>
</tr>
<tr>
<td>Wang et al. [12]</td>
<td>94.86</td>
</tr>
<tr>
<td>Lehrmann et al. [10]</td>
<td>90.90</td>
</tr>
<tr>
<td>Our method</td>
<td>94.89</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Before reduction</th>
<th>After reduction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean F-score</td>
<td>err_{OOB}</td>
</tr>
<tr>
<td>0.94 (+/-0.02)</td>
<td>0.008</td>
</tr>
<tr>
<td>0.94 (+/-0.02)</td>
<td>0.004</td>
</tr>
</tbody>
</table>

![Fig. 4 OOB error rates in terms of RF parameters (n_{trees} and \text{max features})](image4.png)

![Fig. 5 Optimal feature subset with minimal OOB error using Tukey’s test](image5.png)
extraction with LMA approach and after RDF method for actions recognition. We measured the recognition rate using same validation method as MSRC-12 dataset, the 5-fold cross validation technique. We obtained as result the mean of f-scores $0.96(+/-0.02)$. With the same RDF parameters setting in MSRC-12 dataset, we applied features reduction step. Table III summarizes recognition results ($Mean F-score$, OOB error rate ($err_{OOB}$), and number of features ($N$)) before and after applying features reduction step. With 36 features we obtained same mean f-score and a lower OOB error value. So we can say that our method managed to reduce features while keeping most relevant features and same recognition results.

### IV. Conclusion

In this paper, we presented an efficient method for extracting most relevant motion descriptors for human motion recognition. Our descriptor was inspired from LMA technique to combine both quantitative and qualitative characteristics of motion. Furthermore, an effective feature reduction algorithm was applied to keep only the most informative features which had a great impact on the computational latency while maintaining or even improving the reported results. Based on these results, we plan to recognize expressive motions and study the importance of each LMA features to characterize human emotions.

### REFERENCES


