An Efficient Motion Recognition System Based on LMA Technique and a Discrete Hidden Markov Model

Insaf Ajili, Malik Mallem, Jean-Yves Didier

An Efficient Motion Recognition System Based on LMA Technique and a Discrete Hidden Markov Model

Insaf Ajili, Malik Mallem, and Jean-Yves Didier

Abstract—Human motion recognition has been extensively increased in recent years due to its importance in a wide range of applications, such as human-computer interaction, intelligent surveillance, augmented reality, content-based video compression and retrieval, etc. However, it is still regarded as a challenging task especially in realistic scenarios. It can be seen as a general machine learning problem which requires an effective human motion representation and an efficient learning method. In this work, we introduce a novel descriptor based on Laban Movement Analysis technique, a formal and universal language for human movement, to capture both quantitative and qualitative aspects of movement. We use Discrete Hidden Markov Model (DHMM) for training and classification motions. We improve the classification algorithm by proposing two DHMMs for each motion class to process the motion sequence in two different directions, forward and backward. Such modification allows avoiding the misclassification that can happen when recognizing similar motions. Two experiments are conducted. In the first one, we evaluate our method on a public dataset, the Microsoft Research Cambridge-12 Kinect gesture data set (MSRC-12) which is a widely used dataset for evaluating action/gesture recognition methods. In the second experiment, we build a dataset composed of 10 gestures (Introduce yourself, waving, Dance, move, turn left, turn right, stop, sit down, increase velocity, decrease velocity) performed by 20 persons. The evaluation of the system includes testing the efficiency of our descriptor vector based on LMA with basic DHMM method and comparing the recognition results of the modified DHMM with the original one. Experiment results demonstrate that our method outperforms most of existing methods that used the MSRC-12 dataset, and a near perfect classification rate in our dataset.

Keywords—Human Motion Recognition, Motion representation, Laban Movement Analysis, Discrete Hidden Markov Model.

I. Introduction

Human motion recognition is an active area of research due to its importance in many applications: video surveillance, indexing videos, interaction Human-machine, security, and health-care. The purpose of a human motion recognition system is to recognize simple actions of everyday life such as running, knocking, eating, walking, etc.) from videos. The problem of human motions recognition has attracted the attention of several researchers and the advantages and limitations of the different proposed approaches have been discussed over the last years. Two crucial aspects of motion recognition are to extract relevant features by representing the contents to be classified by a descriptor vector and to develop a robust learning algorithm in order to associate with this representation a label. Many descriptors have been chosen in the computer vision literature. First methods based on interest points have been proposed to describe human movements by Laptev and Lindberg [6]. They have used the Harris 3D point-of-interest detector, which is an extension of the Harris detector, adding to it the temporal dimension. Dollar et al. [3] have also proposed a similar detection algorithm, the cuboid detector, based on interests points calculated from Gabor's filter responses in the space and time domain. Other approaches have illustrated the relevance of tracking point trajectories for recognizing actions in videos. For instance, Messing et al. [10] have extracted features trajectories by tracking Harris3D interest points with the help of the KLT tracker. Also Matikainen et al. [9] have extracted trajectories of tracked feature points in a bag of words paradigm for action recognition. Wang et al. [15], have extracted features aligned with the trajectories to characterize appearance and motion. To reduce the influence of camera motion on action recognition, they introduced a descriptor based on motion boundary histograms (MBH) which rely on differential optical flow. When trying to identify human motions, it is sometimes wise to know where the actor of the action is in order to isolate him from the rest of the scene. This allows focussing on his movements, regardless of what happened in the background. To do this, two paths have been widely studied, the first one was the analysis of the shape of the character, through his silhouette, here we can cite the work of Shao and Chen [12] who have employed body poses sampled from silhouettes that were fed into a bag-of-words model. Also, Ahmad and Lee [1] have proposed a spatiotemporal silhouette representation, called Silhouette Energy Image (SEI) to differentiate the properties of form and motion for the human action recognition. The second path was the analysis of the movement of the actor through the identification of his limbs (hands, head, legs, etc.). Zanfir et al. [17] have proposed a moving pose descriptor defined by both pose information and differential information (velocity and acceleration). Hussein et al. [5] have introduced the covariance matrix of skeleton joint locations over time as a descriptor vector. In order to code the temporal dependence of the joints positions, multiple covariance matrices have been deployed over sub-sequences in a hierarchical way. Yang et al. [16] have introduced both spatial and temporal aspects. Their descriptor vector included three pieces of information, the static pose posture information (fcc), the temporal movement of a pose defined by the difference between the current and the previous pose (fcp) and the offset from an initial pose (fci).

I.Ajili, M. Mallem and J. Didier are with IBISC, Univ Evry, Université Paris-Saclay, 91025, Evry, France.

In our work, we tried to propose more suitable features and advanced learning algorithm to improve human motion recognition performance. Our descriptor based on a Laban Movement Analysis method (LMA) introduced by Rudolf Laban (1879 to 1958) to analyze, describe, visualize and annotate all varieties of human motion using a specific notation. LMA is a descriptive language widely used in the field of dance, physical therapy, athletics, and behavioral science. It is generally used to analyze the movement of dancers and athletes. It captures both quantitative and qualitative aspects of the movement by encoding Laban components. To model human motion data we used DHMM method for motions recognition. Classification step is based on Forward algorithm [11] where each motion is presented with two DHMMs to encode motion sequence in the forward and backward directions. By applying Forward algorithm in the two cases we avoid the conflict that can happen between some gestures which share a same part of the motion. For example, extend your arm straight forward and knocking the door are two actions having the same motion in the first frames, which can lead to a misclassification between them. So, the idea here is to classify both motions with taking into consideration the two motion sequence directions which can help to make a distinction between them.

The remainder of the paper is organized as follows. In Section 2, we describe our proposed approach starting with preprocessing data step, feature extraction and finally gesture recognition step. Experimental results are reported in section 3. We first evaluate our descriptor vector with a public dataset MSRC-12 and with our dataset dedicated to control gestures. We compare between recognition results obtained by the basic DHMM method and our proposed DHMM method. Finally, conclusion and future work are presented in Section 4.

II. Human motion recognition system

In general, a human motion recognition system is composed of three important steps: preprocessing data, feature extraction, and motion recognition (Fig.1). In the first step, we introduce an invariant approach to make our system independent of the initial position and orientation of the user. In the following step, to represent human motion, we used three LMA components, Body, Space, and Shape. We didn't use Effort component because it describes the qualitative use of energy and the inner attitude. Effort qualities depend on the rhythm, weight and the intention of the motion and are often used to describe emotions. However, our application consists in recognizing human gestures regardless their rhythm. So, if we make the same gesture at different speeds our system should give the same result. In the recognition step, we used DHMM model which accepts discrete values as input. A discretization approach was implemented following a quantization algorithm (kmeans) in order to generate a set of discrete values which will be implemented into Baum Welch algorithm for training data. In the classification step, we presented each gesture with two DHMM models to process the gesture sequence in two opposite directions and apply the forward algorithm to conclude the action label for a testing gesture.

A. Data Collection And Preprocessing

Prior to the extraction of features, a preprocessing data step is applied to each gesture sequence captured by kinect sensor. So, first we define a local skeleton coordinate system (X', Y', Z') which origin is the hip center joint. $X' - axis$ is the vector starting from the right hip center and going to the left hip center. $Y' - axis$ is the vector connecting between the midpoint of hips (J_c) and the torso joint (P_t). And finally $Z' - axis$ is orthogonal to both vectors.

$$i' = \frac{J_{lhi} - J_{rhi}}{\|J_{lhi} - J_{rhi}\|}; j' = \frac{J_s - J_c}{\|J_s - J_c\|}; k' = i' \wedge j' \quad (1)$$

Let $[J_j]_{(X,Y,Z)}$ be the 3D position of joint j presented in the camera coordinate system (X, Y, Z). We translate the skeleton coordinate system to the center of kinect:

$$[J_j]_X{}' = [J_j]_X - c_x; [J_j]_Y{}' = [J_j]_Y - c_y; [J_j]_Z{}' = [J_j]_Z - c_z \quad (2)$$

Where $J_c(c_x, c_y, c_z)$ is the 3D position of the hip center joint in the kinect coordinate system. After we apply a rotation to align both coordinate systems, we have:

$$[J_j]_{(X,Y,Z)} = R_{(X,Y,Z)\leftarrow(X',Y',Z')}[J_j]_{(X',Y',Z')} \quad (3)$$

$$= [r_1\ r_2\ r_3][J_j]_{(X',Y',Z')} \quad (4)$$

where r_1, r_2 and r_3 are the rotation vectors around the $X - Axis$, $Y - Axis$ and $Z - Axis$ respectively.

By applying (4), we have:

$r_1 = \frac{[J_{lhi}]_{(X,Y,Z)}}{a}$; $r_2 = \frac{[J_s]_{(X,Y,Z)}}{c}$ and $r_3 = r_1 \wedge r_2$ where a is the distance between C and J_{lhi}, and J_c is the distance between J_c and J_s. Since r_1, r_2 and r_3 are orthogonal unit vectors, hence $R^T = R^{-1}$. Then:

$$[J_j]_{(X',Y',Z')} = R^T_{(X,Y,Z)\leftarrow(X',Y',Z')}[J_j]_{(X,Y,Z)} \quad (5)$$

B. Feature Extraction

Feature extraction can be defined as the extraction of significant features from raw data, which maximizes the difference between class patterns while enhancing the variability between class patterns. Our descriptor vector is derived from LMA approach which can describe and interpret all varieties of human movements. For human motion representation we quantify three LMA components: **Body component,** expresses which body parts are moving and how their movement are related to each other. It also addresses issues concerning locomotion and kinematics by describing structural and physical characteristics of the human body. Body organization is related to the connection between body parts. In the upper body part, we describe the extension of elbows $(\theta_1^l, \theta_1^r,)$ and shoulders (θ_2^l, θ_2^r). And we describe outstretched arms by computing the distance between two hands (d_{Hs}). To know more about hands pathway, we add three others features, the distances between the shoulder center and left $(d_{shc,lh})$ and right hand $(d_{shc,rh})$, and the angle between two hands with respect to shoulder

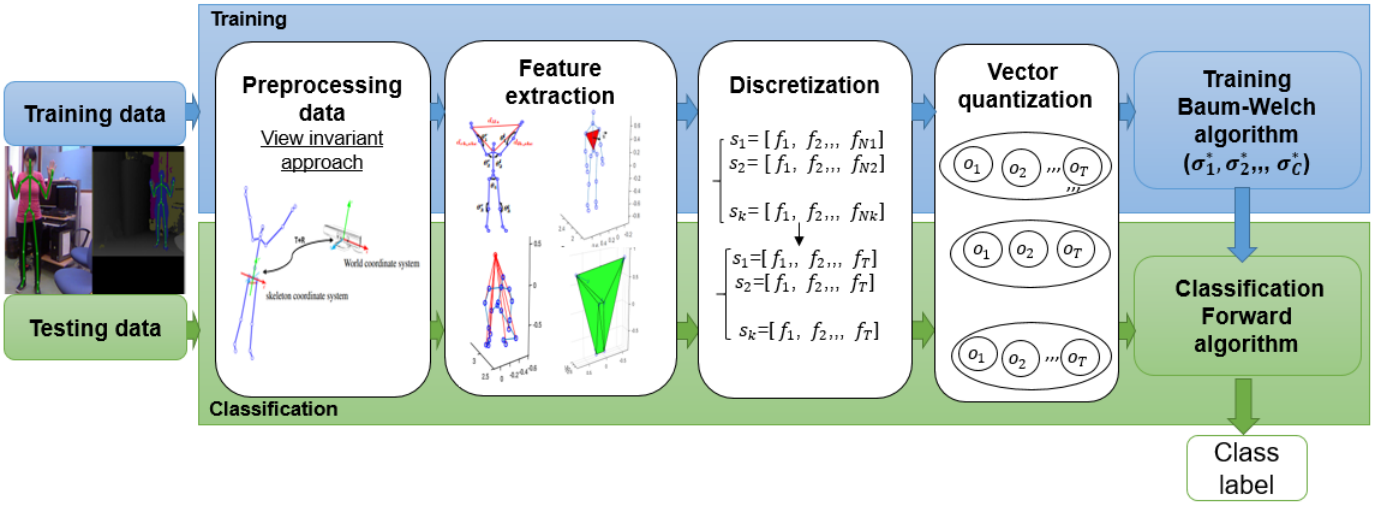


Fig. 1. Human motion recognition steps.

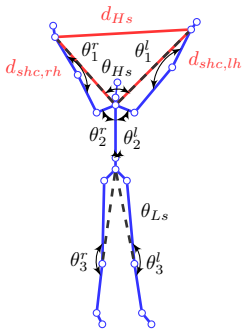


Fig. 2. Body characteristics.

center joint(θ_{Hs}). In the lower part we describe legs extension (θ_3^l, θ_3^r)) and legs spread (θ_{Ls}).

Space component defines where in space the motion is happening, the directions and spatial patterns. We define the torso direction by computing the normal vector $\vec{N}$ of the triangle formed by left hip (hi_l), right hip (hi_r) and neck (n) joints.

$$\overrightarrow{N} = \frac{\overrightarrow{nhi_l} \wedge \overrightarrow{nhi_r}}{||\overrightarrow{nhi_l} \wedge \overrightarrow{nhi_r}||} \tag{6}$$

Shape component describes the way the body changes shape during movement. It focuses on two main qualities, "What forms does the body make"?, "Is the shape changing in a self-to-self relationship or in relation to a goal in space"? ". Shape category is composed of three subcategories: Shape Flow, Carving, and Directional Movement.

Shape flow represents the relationship of the body to itself. It is related to the shape deformation, expanding or condensing, during movement. We compute the volume of the smallest convex envelope of the human body based on Quickhull algorithm [2]. *Directional movement,* represents the pathway through space of the movement, either curving or straight. We quantified how curved are the trajectories made by the upper extremities (head and hands) by computing the gradual angular change ϕ occurring by each joint between two successive frames.

$$\phi_{J_t} = \arccos\left(\frac{\overrightarrow{J_{t-1}J_t}}{\left\|\overrightarrow{J_{t-1}J_t}\right\|} \cdot \frac{\overrightarrow{J_tJ_{t+1}}}{\left\|\overrightarrow{J_tJ_{t+1}}\right\|}\right) \tag{7}$$

where $J_{t/t+1}$ is the position of joints (head and hands) at frame t and $t+1$, respectively. This equation describes the local curvature of the upper body parts'pathway. So in direct movement with straight-line trajectories, local curvature feature will be closest to 0. But in curved trajectories, it will be higher.

Carving describes the qualitative changes in the shape according to three planes horizontal, frontal, and sagittal, and relating them to bipolar descriptors like spreading and enclosing, rising and sinking, and retreating and advancing, respectively. We quantify Carving factor by computing the projected distances between head, upper and lower extremity joints (hands, elbows, knees, and feet) relating to spine joint pose at the initial frame. These projections provide the frontal, sagittal and horizontal displacements of the head, arms, and legs.

$$d = \sqrt{\sum_d (p_{je} - p_{se})^2} \tag{8}$$

where j represents each joint considered at each frame, s is the spine joint at the initial frame and e belongs to one of the following sets $\{x, y\}, \{y, z\}$ and $\{z, x\}$ for each considered projection.

C. Recognition phase

The problem of human motions recognition can be reduced to a problem of supervised classification. In order to make a decision on a given motion, the system performs two steps: training and testing. The purpose of the training phase is to build a set of rules that will be used for the recognition of future actions. Indeed, based on the labeled data (the truth

ground), the system is capable of building decision rules to be able to distinguish between the different categories of human motions. By applying these decision rules to a given action, the system is able to predict its class. In this paper, we use the Hidden Markov Model for both training and classifying motions.

1) Discrete Hidden Markov Model: Hidden Markov Models [11] is one of the most well-known methods in machine learning. An HMM is a statistical model for time series, used to represent the evolution of observable sequences (O) that depend on unobserved, or discrete state variables (S). We modeled each gesture with a left-right HMM. In such model, only transitions from one state to itself or to a unique successor are allowed.

A set of trained HMMs for the C classes of gestures can be represented as $\sigma_{\{1,\dots,C\}} = \{\sigma_1, \sigma_2, \dots, \sigma_C\}$.

The HMMs we use are discrete HMMs and discrete HMMs. They accept only discrete values as inputs. Thus, given our descriptor vector, before implementing it into an HMM, we use discretization and quantization algorithms during training and classification steps. In discretization approach, we implemented a C++ algorithm presented in Algorithm 1 which consists in sampling gesture sequences with different sizes into a fixed-size T. A gesture sequence $s_{in} = \{f_1, f_2, \dots, f_N\}$ is defined as a $N \times d$ matrix composed of N feature vectors with d features, f_k is a feature vector recording at frame k with size d (d=number of features). The output of our discretization algorithm is a $T \times d$ matrix presented a gesture sequence composed of T feature vectors $s_{out} = \{h_1, h_2, \dots, h_T\}$.

Algorithm 1: Discretization algorithm.

Input : $s_{in} = \{f_1, f_2, \dots, f_N\}$, T
Output: $s_{out} = \{h_1, h_2, \dots, h_T\}$

$g_1 = f_1$
$j \leftarrow 2$
$k \leftarrow 1$
for $i \leftarrow 2$ **to** N **do**
  Compute $D = \sqrt{\sum_{l=1}^{d}(f_{i,l} - f_{k,l})^2}$ ▷ D *is the distance between two feature vectors* f_i *and* f_k.
  if $D \geq \epsilon$ **then**
    $g_j = f_i$
    $j \leftarrow j + 1$
    $k = i$
  end
end
Compute the average distance: $D' = \frac{N'}{T}$ ▷ g *is the matrix of* N' *feature vectors obtained after removing noise.*
for $i \leftarrow 0$ **to** $T - 1$ **do**
  $h_{i+1} = g_{1+i*D'}$.
end

After the discretization step, we apply k-means algorithm [8] to cluster the feature vectors of all gesture sequences into K clusters $\{c_1, c_2, \dots, c_K\}$ in which each feature vector belongs to the closest cluster, so as to satisfy the condition expressed by:

$$\underset{c}{argmin} \sum_{j=1}^{K} \sum_{h_i \in c_j} \|h_i - \mu_j\|^2 \tag{9}$$

where μ_j is the mean of the elements in the cluster c_j. At the end of the quantization algorithm, a gesture sequence will be presented as an observation sequence $O = \{o_1, o_2, \dots, o_T\}$, where o_i is a discrete symbol $\in \{c_1, c_2, \dots, c_k\}$. So each symbol o_i corresponds to the cluster of the feature vector f_i, and the output sequence length is T, which acts as the discrete observation sequence to be the input observation sequences to learn the HMM model and use this model to predict for the unknown sequence.

The parameters of the model HMM can be represented in the compact way $\sigma = (\pi, A, B)$, where π is the initial probability distribution over states, A is the transition probability matrix, and B is the matrix that represents the emission probability of a symbol observed from a specific state.

For the training step, we use the Baum-Welch algorithm to find optimal parameters to the HMM given an initial model $\sigma_i = (\pi_i, A_i, B_i)$ and the observations sequences $\{O_1, O_2, \dots, O_s\}$ corresponding to the learning gesture sequences.

$$\sigma^* = \underset{\sigma}{argmax}(\sum_{i=1}^{s} logP(O_i|\sigma)) \tag{10}$$

For classification step, we use the Forward algorithm to classify a sequence test $O_t = \{o_1, o_2, \dots, o_T\}$. The class label is assigned via Maximum Likelihood after evaluating the sequence in every HMM.

$$C = \underset{all\sigma}{argmax}(log(P(O_t|\sigma)) \tag{11}$$

2) Modified DHMM: In the classification step, sometimes we have very similar motions, in such case our recognition algorithm can make a misclassification gesture. An example is provided in Fig.3 which shows two similar gestures ($G1$ and $G2$). Gesture 2 shares a part with Gesture 1 in the first three initial frames. In such case, initial state probabilities and transition probability matrices of the two models are very close which can lead to an error classification.

The idea here is to process the gesture sequence in two directions the forward (from the first frame to the last frame) and backward (from the last frame to the first frame). We define two models for each gesture σ_G^d and σ_G^i are the DHMM models when considering the gesture sequence (G) in the forward direction and in the backward direction, respectively. If we take a test sequence (O_t^d) of the second gesture ($G2$), we have:

$$P(O_t^d|\sigma_{G_1}^d) \approx P(O_t^d|\sigma_{G_2}^d) \tag{12}$$

Now if we consider the observation sequence in the backward direction (O_t^i) we have:

$$P(O_t^d|\sigma_{G_2}^d) \approx P(O_t^i|\sigma_{G_2}^i) \tag{13}$$

$$P(O_t^i|\sigma_{G_2}^i) > P(O_t^i|\sigma_{G_1}^i) \tag{14}$$

According to (12)_(14), we obtain the following Equation:

$$min(P(O_t^d|\sigma_{G_2}^d), P(O_t^i|\sigma_{G_2}^i)) > min(P(O_t^d|\sigma_{G_1}^d), P(O_t^i|\sigma_{G_1}^i)) \quad (15)$$

According to (15), we can declare the class label of the test sequence (O_t) as:

$$C = \underset{all\sigma}{argmax}\, min(log(P(O_t^d|\sigma_G^d)), log(P(O_t^i|\sigma_G^i))) \quad (16)$$

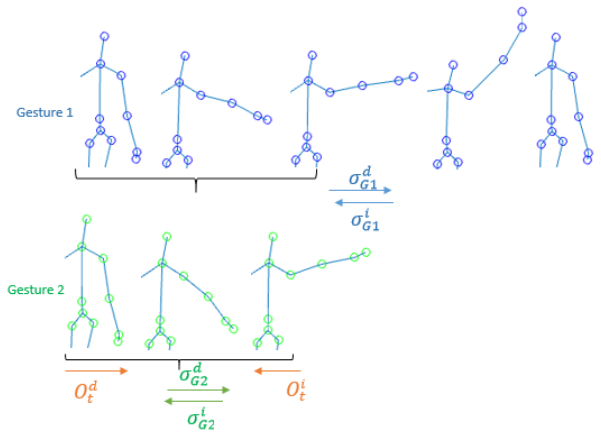


Fig. 3. Processing gestures sequences in two opposite directions for two similar gestures.

III. EXPERIMENTAL RESULTS

We evaluated the robustness of our descriptor and our classification method for action recognition. We performed this evaluation on two datasets: The first one is a public dataset (MSRC-12) acquired using a Kinect sensor and the second dataset is our dataset dedicated to control gestures built under Robot Operating System (ROS). We applied our descriptor vector and modified DHMM method for action recognition step. Details of the experiments are presented in the following subsections.

A. MSRC-12 Dataset [4]

To evaluate our approach, we tested our method on a relatively large dataset captured by a Kinect sensor, the MSRC-12. The dataset is composed of 12 gestures performed by 30 subjects. Each subject repeats the same gesture several times. In total, there are 6244 gesture instances. The motion files contain 3D coordinates of 20 joints captured at a sample rate of 30Hz with an accuracy of about 10 centimeters in joint positions.

The dataset is divided into two groups: iconic (hide, shoot pistol, throw object, change weapon, kick, put goggles) and metaphoric (raise volume, navigate to next menu, wind up music, take a bow, protest music, low down song) gestures. We converted the raw data into a descriptor vector based on the three LMA components. After, we discretized gesture sequences into a fixed length by applying discretization method presented in Algorithm 1 with $T = 70$ frames. A grid search on DHMM parameters (number of states S, and number of symbols O) ranging from 5 to 40 has been done. Best results are achieved with $S = 20$ states and $O = 40$ symbols.

TABLE I
COMPARISONS WITH STATE-OF-THE-ART APPROACHES ON THE MSRC-12 DATASET.

Methods	Iconic	Metaphoric
Lehrmann et al. [7]	90.90	-
Song et al. [13]	79,77	81
Truong et al. [14]	88.6	75.2
Ours (DHMM)	96.33	90.66

In order to compare our approach with state of the art methods, we used the more challenging validation method, the cross-subject test where one-half of the subjects are used for training and the remaining for testing.

Fig.4 demonstrates that our method achieves an average accuracy of 94.03% and 80.48% on iconic and metaphoric gestures respectively when applying basic DHMM method. With our modified DHMM we improved the recognition results with an average accuracy of 96.33% for iconic gestures and 90.66% for metaphoric gestures.

The results in Table I show the comparison between our result and state of art results. For a faithful comparison, we take the result of Truong et al. [14] where they used the same learning method DHMM. Our method outperforms their method by 7.73% on iconic gestures and 15.46% on metaphoric gestures. In general, we can say that our method performs better than the state-of-the-art result.

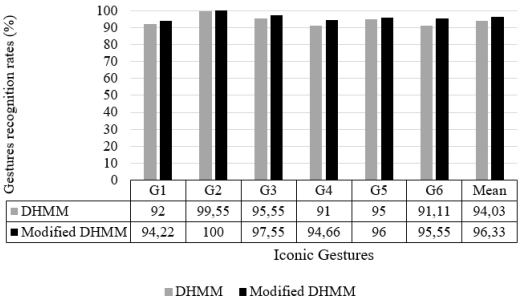


Fig. 4. Recognition rates results of MSRC-12 iconic gestures when applying basic DHMM and modified DHMM.

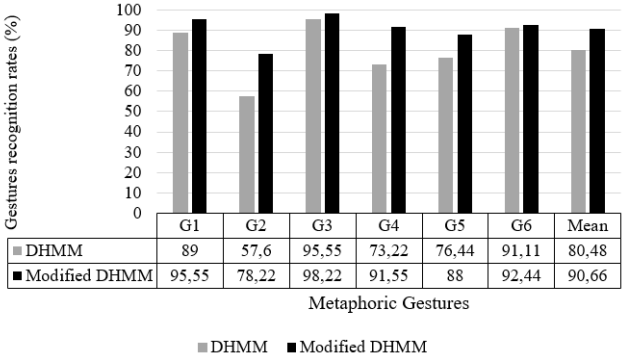


Fig. 5. Recognition rates results of MSRC-12 metaphoric gestures when applying basic DHMM and modified DHMM.

B. Control gestures dataset

After evaluating our system with a public dataset, we built our dataset composed of ten control gestures (move, introduce, turn left, stop, turn right, increase velocity, decrease velocity, waving, dance, introduce yourself) as shown in Fig.9.

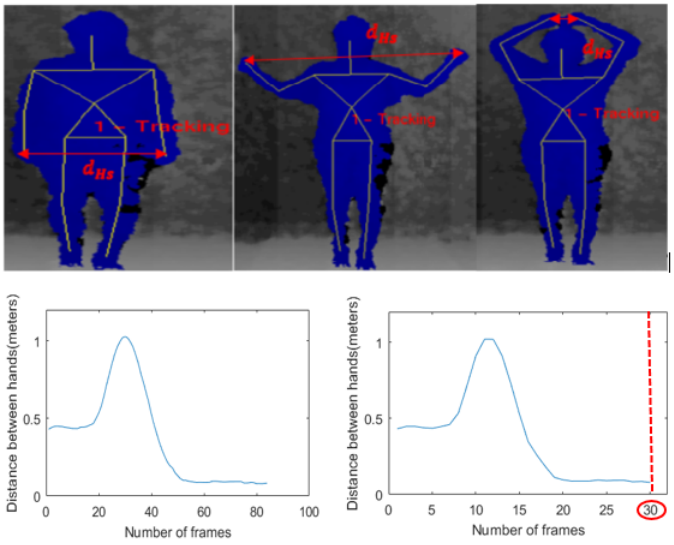


Fig. 6. Discretization of d_{Hs} feature in "Stop" gesture with $T = 30$ frames.

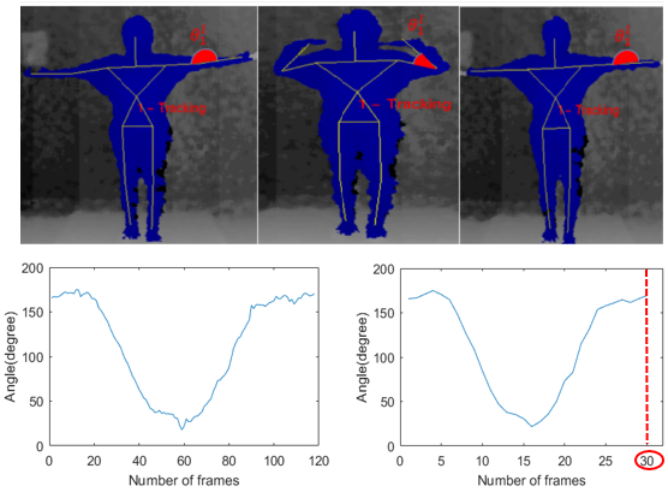
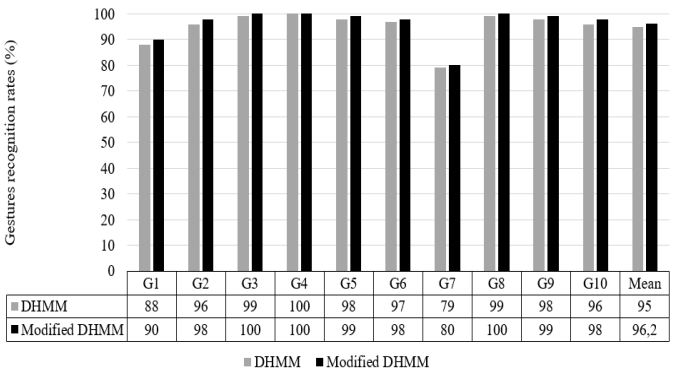


Fig. 7. Discretization of θ_1^l feature in "Waving" gesture with $T = 30$ frames.

Twenty subjects (10 men and 10 women) from the University of Evry Val d'Essonne, ranged in age from 27 to 45 years old (M=28.5 years, SD=5.5) took part in this study. Each subject is asked to make gesture ten times. Our dataset has in total 2000 sequences (20 subjects$\times$ 10 gestures$\times$ 10 times). Our system was implemented under ROS, a Robot Operating System which consists in running a great number of executables to exchange data synchronously (via topics) or asynchronously (via services). For data acquisition, the OpenNI driver provided a high-level skeleton tracking module. This module requires initial calibration to record the 3D position of skeleton joints at 640×480 resolution at 30 fps. For each gesture, we computed the descriptor vectors from raw data. After, we sampled descriptor vectors with 30 frames. Two examples of the discretization of two extracted features from our descriptor vector are shown in Fig.6 and Fig.7. The first feature (d_{Hs}) is the distance between hands in "stop" gesture. As we can see in Fig.6, when performing stop gesture this feature tends to zero at the end of the gesture. When applying our discretization algorithm with T= 30 frames, we convert the gesture sequence from a vector with 84 frames to a vector with 30 frames. For the "waving" gesture we take as feature example the angle between left hand and left shoulder. As shown in Fig.7, at initial frame the angle $\theta_2^l \approx 180\,\text{deg}$ after it drops to $\approx 20\,\text{deg}$ and finally it returns to its initial value ($\approx 180\,\text{deg}$). We sampled the feature sequence of θ_2^l in "waving" gesture from 118 frames to 30 frames. We set the number of clusters in K-means algorithm to 20 and the number of hidden states to 5. As shown in Fig.8, our results were very satisfying with an average accuracy result of 95% when applying simple DHMM and a significant result of 96.2% with our modified DHMM.



	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10	Mean
DHMM	88	96	99	100	98	97	79	99	98	96	95
Modified DHMM	90	98	100	100	99	98	80	100	99	98	96,2

Fig. 8. Recognition rates results of our dataset when applying basic DHMM and modified DHMM.

IV. Conclusion

In this paper we presented an effective approach for human motion recognition based on specific feature vectors inspired from LMA technique. A series of steps have been implemented starting with preprocessing data step based on view-invariant human motion, folowing by a suitable feature extraction method, ending with a robust recognition method based on Discrete Hidden Markov Model in order to improve the recognition accuracy of our approach. Experimental results on MSRC-12 dataset show that our method proves to be superior to some state of the art methods for skeleton-based recognition. A perfect classification performance was achieved in our dataset composed of ten control gestures. The future work will focus on enhancing our dataset by introducing expressive gestures, means gestures performed with different emotions. The main purpose of this idea is to perform expressive communicative gestures for a Human-Robot interaction.

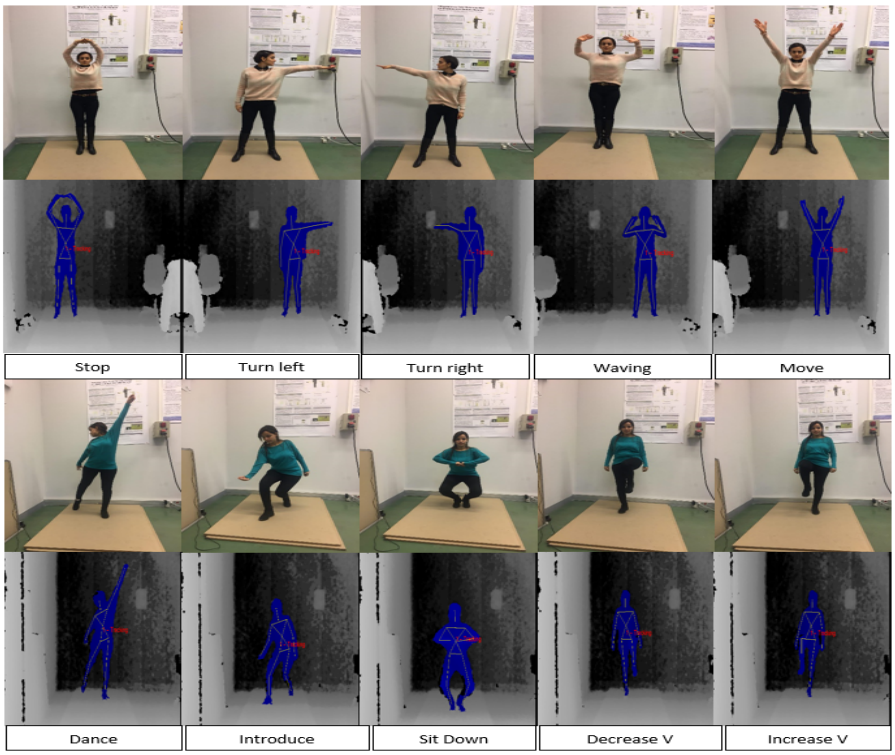


Fig. 9. Control gestures dataset.

REFERENCES

[1] M. Ahmad and S.-W. Lee. Variable silhouette energy image representations for recognizing human actions. *Image and Vision Computing*, 28(5):814 – 824, 2010. Best of Automatic Face and Gesture Recognition 2008.

[2] C. B. Barber, D. P. Dobkin, and H. Huhdanpaa. The quickhull algorithm for convex hulls. *ACM Trans. Math. Softw.*, 22(4):469–483, Dec. 1996.

[3] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *2005 IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, pages 65–72, Oct 2005.

[4] S. Fothergill, H. Mentis, P. Kohli, and S. Nowozin. Instructing people for training gestural interactive systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, pages 1737–1746, New York, NY, USA, 2012. ACM.

[5] M. E. Hussein, M. Torki, M. A. Gowayyed, and M. El-Saban. Human action recognition using a temporal hierarchy of covariance descriptors on 3d joint locations. In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*, IJCAI '13, pages 2466–2472. AAAI Press, 2013.

[6] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, June 2008.

[7] A. M. Lehrmann, P. V. Gehler, and S. Nowozin. Efficient nonlinear markov models for human motion. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1314–1321, June 2014.

[8] J. Macqueen. Some methods for classification and analysis of multivariate observations. In *In 5-th Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297, 1967.

[9] P. Matikainen, M. Hebert, and R. Sukthankar. Trajectons: Action recognition through the motion analysis of tracked features. In *2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops*, pages 514–521, Sept 2009.

[10] R. Messing, C. Pal, and H. Kautz. Activity recognition using the velocity histories of tracked keypoints. In *2009 IEEE 12th International Conference on Computer Vision*, pages 104–111, Sept 2009.

[11] L. R. Rabiner and B. H. Juang. An introduction to hidden Markov models. *IEEE ASSP Magazine*, pages 4–15, January 1986.

[12] L. Shao and X. Chen. Histogram of body poses and spectral regression discriminant analysis for human action categorization. In *BMVC*, 2010.

[13] Y. Song, L. P. Morency, and R. Davis. Distribution-sensitive learning for imbalanced datasets. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–6, April 2013.

[14] A. Truong and T. Zaharia. Dynamic gesture recognition with laban movement analysis and hidden markov models. In *Proceedings of the 33rd Computer Graphics International*, CGI '16, pages 21–24, New York, NY, USA, 2016. ACM.

[15] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu. Dense trajectories and motion boundary descriptors for action recognition. *International Journal of Computer Vision*, 103(1):60–79, May 2013.

[16] X. Yang and Y. Tian. Effective 3d action recognition using eigenjoints. *Journal of Visual Communication and Image Representation*, 25(1):2 – 11, 2014. Visual Understanding and Applications with RGB-D Cameras.

[17] M. Zanfir, M. Leordeanu, and C. Sminchisescu. The moving pose: An efficient 3d kinematics descriptor for low-latency action recognition and detection. In *2013 IEEE International Conference on Computer Vision*, pages 2752–2759, Dec 2013.