Latent Force Models for Human Action Recognition

Zhi Chao Li, Ryad Chellali(✉), and Yi Yang(✉)

CEECS- Nanjing Robotics Institute, Nanjing Tech University,
Nanjing, People's Republic of China
{rchellali,yyang}@njtech.edu.cn

Abstract. Human action recognition is a key process for robots when targeting natural and effective interactions with humans. Such systems need solving the challenging task of designing robust algorithms handling intra and inter-personal variability: for a given action, people do never reproduce the same movements, preventing from having stable and reliable models for recognition. In our work, we use the latent force model (LFM [2]) to introduce mechanistic criteria in explaining the time series describing human actions in terms actual forces. According to LFM's, the human body can be seen as a dynamic system driven by latent forces. In addition, the hidden structure of these forces can be captured through Gaussian processes (GP) modeling. Accordingly, regression processes are able to give suitable models for both classification and prediction. We applied this formalism to daily life actions recognition and tested it successfully on a collection of real activities. The obtained results show the effectiveness of the approach. We discuss also our future developments in addressing intention recognition, which can be seen as the early detection facet of human activities recognition.

Keywords: Skeleton model · Latent force model · Gaussian processes regression · Feature modeling

1 Introduction

Human actions recognition is getting more attention the last years. This task is of interest in many fields such as robotics, computer vision, human-computer interaction, and natural language processing, etc. targeting applications in homecare, personal and manufacturing robotics, behavior analysis and many other domains. In our case, we are interested in human-robots interactions (HRI), where robots need to understand the actual contexts to better serve humans. Human actions are an important part of these contexts and actions recognition capability is a key feature for friendly and accepted personal robotics.

A lot of works have been done in HAR using videos [9]. The developed techniques use image sequences (2D information evolving in time) and analyze the spatiotemporal changes of human bodies appearance to infer human actions or activities. Recently, RGB-D cameras were introduced. This allows using more reliable data as they encode time series describing human postures and skeletons with avoiding classical issues such

N. Kubota et al. (Eds.): ICIRA 2016, Part I, LNAI 9834, pp. 237–246, 2016.
DOI: 10.1007/978-3-319-43506-0_20

as occlusions, view point changes, lightning, etc. In its generality, the human actions recognition problem can be seen as a general pattern recognition problem: basically, one needs to match the observation, a multivariate time series to a previously seen pattern and assign a label to it, i.e. an action.

Most of existing skeleton based techniques model explicitly the temporal dynamics of skeleton joints. The dynamics are expressed as local models such as ARMA in [10] or sequential state transition models (graphical models) such as HMM or DBN [11, 12].

In our work, we consider the human body as a dynamic system and we describe the dynamics through latent force models (LFM [2]) for which, the human body is seen as a dynamic system driven by latent forces. The LFM has been used mainly for prediction purposes. Here, we adapt it to handle recognition tasks. The LFM two main advantages: (1) as other dynamic systems based techniques, it allows understanding the skeleton times series as HA, (2) it introduces a mechanistic flavor, which can be advantageously used to enhance both the robustness and the interpretability of the sequences. Indeed, observed body movements are generated by latent forces (muscles) and any mechanistic model could be mapped to these forces even indirectly (the general case of LFM).

In the following, we explicit the LFM model and the way we derive it from the raw skeleton data. In Sect. 4, we show how the obtained LFM model is used for classification purposes. Mainly we will demonstrate how to combine the forces and the sensitivity parameters (the weights of the forces) in order the feed a simple linear SVM to perform the classification task. Finally, we present our experimental protocol including the used datasets (the MAD dataset from CMU and the daily action dataset we collected) as well as the evaluation methods. We finish by presenting the obtained results and discuss our future works.

2 Related Work

Action recognition research is very active since a decade. Pushed mainly by social networks industry, many impressive researches were achieved based on the analysis of 2D videos (see [9] for a good review). More recently and with the RGB-D cameras, the HAR issue changed slightly. Indeed, in RGB-D videos, accurate depth and skeleton [1] information are available. This simplifies the HRA by removing occlusions and viewpoint related ambiguities (perspective distortion, lack of Euclidian metrics) while providing absolute measurements. In addition, one can use two different cues: depth maps, which can be processed as normal RGB maps and skeletons joints positions, which may reduce the effects of single inputs.

In RGB-D/skeleton data, the geometry is exact. Taking this advantage, some authors proposed encoding HA according to some Euclidian groups formalisms: In [4], the Euclidian rotation-translation group SE(3) is mapped on the Lie group to have compact joints trajectories. The mapped trajectories are then warped with a DTW and a linear SVM are sufficient to perform HRA. More classical is the work in [8]. Authors used classical tools for time series to classify human actions with LDA classifiers.

Grammatical models have been used in [3]: “Bags of words” were constructed from local descriptors to generatively derive *actionlets*, which are considered as the atomic components of actions. Probabilistic graphical models were in fact the most used in HAR. They allow capturing the dynamics of the body with handling the difficult issues of inter and intra-personal discrepancies. In [4], deep belief networks and HMM are combined together to perform simultaneous segmentation and recognition. Piyathilaka et al. [6] developed Gaussian mixture-based HMM for activity recognition [6]. Jaeyong Sung used a hierarchical maximum entropy Markov model detect human actions [7]. In [5], authors present a method of learning latent structures in human actions. Their graphical model relies not only on the body dynamics but also on objects with which humans interact.

Linear dynamic systems (LDS) have been also used. In [11], the HA time series are modeled as an ARMA process and then projected on Grassman manifold, which allows clustering actions.

More recently, hierarchical recurrent neural networks were used to avoid engineering low level coding (i.e. features design) [12]. In their work, authors grouped body parts into subgroups, facilitating the learning process as well as taking advantage of the actual cross-correlation of the subparts when performing actions.

Different from previous works, we take inspiration from the findings of Laurence et al. [2] concerning probabilistic dynamic systems. We extend the latent force model (LFM) to use it for action recognition purposes.

The LFM describes human movements as a dynamic system for which the model can be derived through Gaussian processes regression. This model makes the assumption that latent forces generate body movements, i.e., the forces excite the body to generate the observed data. In the initial formulation, the LFM was used to predict the future body postures/movements. In our work, we use the same formalism with the inclusion of the full set of parameters (the forces and their relative weights) to perform a discriminative recognition using a simple linear SVM. We demonstrate that this formalism is efficient in HAR but not only. Indeed, our approach captures HA per se but also can give mechanistic hints about the movements, which can be of interest in explaining some aspects related to motor activity analysis, as it will be discussed in the conclusion.

3 Action Pattern Representation

This part focuses on the presentation of action and features used for recognizing. In the original work of Alvarez et al., the purpose of using LFM was forecasting. That is to say, given a time series at time t, the LFM is used to predict to future states of system. We modified the formalism in order to use it as a classification framework.

3.1 Action Definition

An action is a physical activity, where ones body perform a set of movements to achieve a predetermined goal such moving an object, changing posture, etc. Formally and considering skeleton based description, an action could be described as a time series of an N-vector, starting at time t_{start} and finishing at time t_{end}:

$$Y(t) = [y_1(t)\, y_2(t) \ldots y_N(t)]$$
$$t \in [t_{start}, t_{end}]$$

Where the y_i are the R joints in the skeleton issued from the RGB-D sensor (Fig. 1).

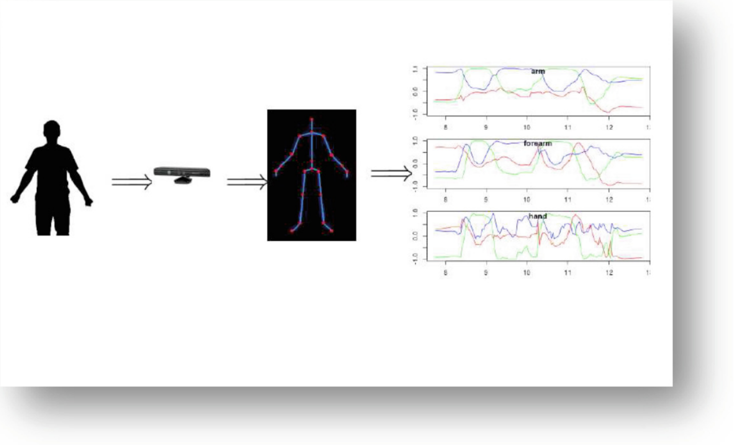


Fig. 1. Body postures time series

3.2 Latent Force Model

In Alvarez et al. introduced latent force models. Basically, the proposed method performs dimension reduction in time series.

$$Y = F.W^T + E$$

Under the assumption of normally distributed noise, the reduced representation or the latent structure F can be seen as a Gaussian process (GP):

$$p(F_r(t)|t;\theta) \sim \prod_{r=1}^{R} N\left(f_r(t)|0, K_r(t,t')\right)$$

Where the posterior F can be derived from the covariance matrix K.

This formulation is then extended to describe the time series as series issued by a dynamic system equivalent to a second order system: a mass spring-damper system excited by latent forces.

$$FS = \ddot{Y}M + \dot{Y} + YD + \sum$$

Where F is the forces matrix, M and C are the mass diagonal mass and damping matrices, D is the original system matrix. These hybrid models consider the human body in movement as a dynamic system driven by some latent forces, non-exactly related to the actual forces (i.e. muscles activity) but allowing involving interesting mechanistic principles to model body parts movements. This model has some similarities with the dynamic movements primitives, which is used in robotics to encode robots trajectories for learning actions by imitation.

Rewriting the differential Eq. (1), we have:

$$\frac{d^2y_i(t)}{dt^2} + C_i\frac{dy_i(t)}{dt} + D_iy_i(t) = \sum_{r=1}^{R} S_{d,r}f_r(t) \tag{1}$$

Where every observed time series $y_i(t)$ *is related to* the R driving $f_r(t)$ latent forces and the N*R constants $S_{n,r}$ sensitivities. The Knowledge of the latent forces, the sensitivities and the constants C and D, it possible to derive the dynamics of the system and to predict its outputs $y_i(t)$. This has been done in [4].

Assuming the latent forces as R independent GPs, it is shown that recovering new values $y_i^*(t)$ is possible through a Gaussian process regression. Indeed, the output covariance can be expressed as linear functions of the latent forces and the covariance of a general stochastic process. From an original sequence $y_i = \{(y_k, t_k)k = 1, \ldots, n\}$, it is possible to predict new values y^* in the sequence. It is shown that $[y,y^*]^{\mathrm{T}}$ satisfies the following distribution:

$$\begin{bmatrix} \boldsymbol{y} \\ \mathbf{y}^* \end{bmatrix} \sim N\left(0, \begin{bmatrix} \mathbf{K}_{yy} & \mathbf{K}_{yy*} \\ \mathbf{K}_{y*y} & \mathbf{K}_{y*y*} \end{bmatrix}\right)$$

Where:

$$\begin{aligned} \boldsymbol{y}^* &\sim N(\mu^*, \sigma^*), \\ \boldsymbol{\mu}^* &= \mathbf{K}_{\boldsymbol{y}}^T \cdot \mathbf{K}_{\boldsymbol{yy}}^T \boldsymbol{y} \\ \sigma^* &= -\mathbf{K}_{y^*}^T \mathbf{K}_{yy}^T \mathbf{K}_{y^*} + \mathbf{K}_{y*y*} \end{aligned}$$

With:

$$K_{y_iy_j}(t, t') = \frac{\sum_{1}^{R} S_{ri}.S_{rj}\sqrt{\pi L_r^2}}{8\omega_i.\omega_j} K_{y_iy_j}^r(t, t'), \omega_p = \sqrt{4.D_p - C_p^2}$$

4 Action Recognition

In our case, the aim is to use the model for classification purposes. In other words, we use the vector $\theta = [\, C_i \quad D_i \quad S_{ir} \quad f_r \quad L_r \,]_{i=1,N;r=1,R}$ to encode actions and use it in a discriminative procedure to perform action recognition.

4.1 Hyper-parameters Learning

In order to derive the model's hyper-parameters, one needs to maximize the following logarithm marginal likelihood function:

$$\log(p(Y|F,\theta)) = -\frac{1}{2}Y^T K_Y^{-1} Y - \frac{1}{2}\log|K_y| - \frac{n}{2}\log 2\pi \tag{2}$$

$K_y = K_{f_r} + \sigma_n^2 I$ is the covariance matrix for noisy inputs Y. K_{fr} is the latent forces covariance without noise, while $(-0.5.Y^T.K_y^{-1}Y)$ expresses the predictions errors, *(–0.5.log*$|K_Y|$) is a penalty depending only on the covariance function and the inputs. $-\frac{n}{2}\log(2.\pi)$ is a normalization constant.

To optimize (2), we use a gradient descent over the vector

$$\theta = [\, C_i \quad D_i \quad S_{ir} \quad f_r \,]_{i=1,N;r=1,R} \tag{3}$$

4.2 Implementation

The pseudo-code of our implementation is the following:

```
Step 1:
Given the sequence $Y = [y_1(t) \quad y_2(t) \quad \ldots \quad y_N(t)]$
Step 2:
Initialize $\theta = [C_i \quad D_i \quad S_{ir} \quad f_r \quad L_r]_{i=1,N;r=1,R}$
Step 3:
For i=1,N
          For j=2,N

$K_{y_i y_j}(t,t') = \frac{\sum_{1}^{R} S_{ri}.S_{rj}\sqrt{\pi L_r^2}}{8\omega_i.\omega_j} K_{y_i y_j}^{r}(t,t')$
          end
end
```

```
Step 4:
With $\theta = [C_i \quad D_i \quad S_{ir} \qquad L_r]_{i=1,N;r=1,R}$
$J^{new} = -\frac{1}{2} Y^T .K_{Y'}^{-1}.Y - \frac{1}{2}\log|K_{Y'}| - \frac{n}{2}\log(2.\pi)$
$K_{y'} = K_y + \sigma^2 .I$, $\sigma$ is a noise
  $if |J^{new} - J^{old}| < \varepsilon$
  $else$
$\theta^{new} = \theta^{old} - \alpha.\frac{\partial J(\theta)}{\partial \theta}$, goto Step 3
Step 5:
For i=1,R
        For j=1,N

$K_{f_r y_j} = \frac{L_r}{q.4.\omega_j}\left[\gamma_r\left(\bar{\gamma},t,t'\right) - \gamma_j\left(\bar{\gamma},t,t'\right)\right]$
        end
end
Step 6:
$F = K_{fy}.K_{yy}.Y$
```

For more readability, we omitted detailing the γ functions (can be found in [2]).

In our case, we considered 13 joints ($N = 13$) for the skeleton and we used two latent forces ($R = 2$). Accordingly, the hyper-parameters vector components are the following:

$$\begin{cases} F = f_{1,2}(t) \\ C = C_{1:13} \\ D = D_{1:13} \\ S = S_{1,2:1,3} \end{cases}$$

4.3 Features Vector and Action Classification

For action recognition, we started considering only the latent forces. That is to say the time series $[f_{1,2}(t)]$. Unfortunately, this information was not sufficient to discriminate among actions. Mainly, we found that similar movements but performed by different body parts, generates similar forces, e.g., raising a hand and raising a leg. On the contrary, the constants of the system, namely, the sensitivities, the damping and the friction parameters were clearly different. This corresponds to the initial intuition that the sensitivities modulate/guide the energy towards some body parts, while the damping/friction represents the sharpness of the observed movement. Accordingly, we constructed our features vector as the concatenation of the forces time series together with the system constants.

$$X = [\, f_{1,2}(t) \quad S_{1,:1,3} \quad S_{1,:1,3} \quad D_{1,13} \quad C_{1,13} \quad]$$

This vector has been used to feed a Support Vector Machine (SVM) to perform the classification.

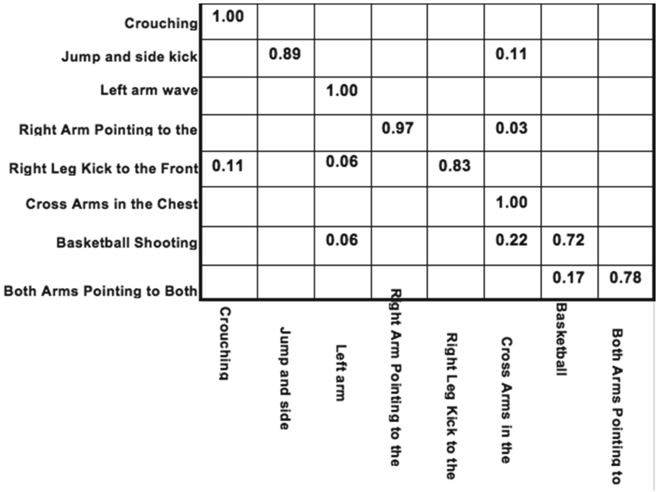
5 Results

Given human activity sequences, action recognition problem consists in solving two sub-problems: (1) segmenting the sequence into actions, (2) recognizing the segmented action. In a previous work [13, 15], the segmentation was addressed and is not considered here and we only focus on segmented sequences.

We evaluate our proposed Latent force-based features on two different datasets. One is the MAD dataset from CMU [14]. The second dataset includes daily actions we collected with a Kinect sensor.

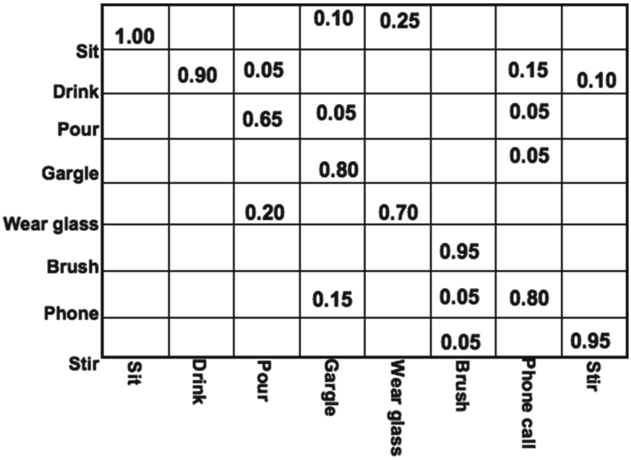
To check the preliminary feasibility, we considered 8 different actions from both datasets. 20 people perform every single action in the MAD dataset twice. We extract latent force features using 13 original joints time series. The homemade dataset is very similar to the MAD. We focused on daily actions, such as drinking, wearing glass and stirring etc. Every action has been performed 20 times from 4 different people. Here as well, every sample has 13 features sequences also joint angles.

Table. 1. MAD dataset results: av precision 89.93 %



	Crouching	Jump and side	Left arm	Right Arm Pointing to the	Right Leg Kick to the	Cross Arms in the	Basketball	Both Arms Pointing to
Crouching	1.00							
Jump and side kick		0.89				0.11		
Left arm wave			1.00					
Right Arm Pointing to the				0.97		0.03		
Right Leg Kick to the Front	0.11		0.06		0.83			
Cross Arms in the Chest						1.00		
Basketball Shooting			0.06			0.22	0.72	
Both Arms Pointing to Both							0.17	0.78

Table 2. The homemade dataset. Av precision 84.37 %



	Sit	Drink	Pour	Gargle	Wear glass	Brush	Phone call	Stir
Sit	1.00			0.10	0.25			
Drink		0.90	0.05				0.15	0.10
Pour			0.65	0.05			0.05	
Gargle				0.80			0.05	
Wear glass			0.20		0.70			
Brush						0.95		
Phone				0.15		0.05	0.80	
Stir						0.05		0.95

Every sample is performed randomly and with different length. The body joints are captured at sample rate 30 Hz but down-sampled to 10 Hz.

After extracting the features in both datasets, we applied the Leave-one-out cross-validation is performed. Table 1, and Table 2 show results, respectively on MAD and on the homemade daily actions. The average precisions are resp. 89.93 % and 84.375 %, which comparable to state of the art results.

6 Conclusion and Future Work

In this paper, we presented latent force features based approach in solving human action recognition. The proposed features reach interesting results compared to existing works. Moreover, it allows more basic interpretation in relation with energetic and biomechanics aspects of human motion. In the near future, our aim is to include these two categories in the analysis to go deeper in human movement interpretation. Indeed, some preliminary tests showed the effectiveness of this coding to interpret some motor impairment such as trembling of limbs. The other point we want to address concerns segmentation. Our previous work will be combined to this one to provide a complete system.

Though the features are good to describe some actions, some actions are not described well by proposed features. A way to improve is to increase the number of latent forces. Unfortunately, this has a computational cost (inverting a large co-variance matrix) and a more adapted optimization technique should be investigated: we used a classical gradient descent while we have a quadratic form and local techniques should be faster avoiding the matrix inversion issue.

References

1. Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., Blake, A.: Real-time human pose recognitionin parts from single depth images. In: CVPR (2011)
2. Alvarez, M.A., Luengo, D., Lawrence, N.D.: Latent force models. In: van Dyk, D., Welling, M. (eds.) Proceedings of 12th International Conference Artificial Intelligence and Statistics, pp. 9–16, April 2009
3. Wang, J., Liu, Z., Wu, Y., et al.: Mining actionlet ensemble for action recognition with depth cameras. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1290–1297. IEEE (2012)
4. Vemulapalli, R., Arrate, F., Chellappa, R.: 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 (2014)
5. Koppula, H.S., Gupta, R., Saxena, A.: Learning human activities and object affordances from RGB-D videos. Int. J. Robot. Res. **32**(8), 951–970 (2013)
6. Piyathilaka, L., Kodagoda, S.: Gaussian mixture based HMM for human daily activity recognition using 3D skeleton features. In: 2013 8th IEEE Conference on Industrial Electronics and Applications (ICIEA), pp. 567–572. IEEE (2013)
7. Sung, J., Ponce, C., Selman, B., et al.: Unstructured human activity detection from RGBD images. In: 2012 IEEE International Conference on Robotics and Automation (ICRA), pp. 842–849. IEEE (2012)
8. Zhang, H., Parker, L.E.: 4-dimensional local spatio-temporal features for human activity recognition. In: 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 2044–2049. IEEE (2011)
9. Cheng, G., Wan, Y., Saudagar, A.N., Namuduri, K., Buckles, B.P.: Advances in Human Action Recognition: A Survey. CoRR abs/1501.05964 (2015). http://arxiv.org/abs/1501.05964
10. Chaudhry, R., Ofli, F., Kurillo, G., Bajcsy, R., Vidal, R.: Bio-inspired dynamic 3D discriminative skeletal features for human action recognition. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops (2013)
11. Slama, R., Wannous, H., Daoudi, M., Srivastava, A.: Accurate 3D action recognition using learning on the Grassmann manifold. Pattern Recogn. **48**(2), 556–567 (2015). Elsevier
12. Du, Y., Wang, W., Wang, L.: Hierarchical recurrent neural network for skeleton based action recognition. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, pp. 1110–1118 (2015). doi:10.1109/CVPR.2015.7298714
13. Bernier, E., Chellali, R., Thouvenin, I.M.: Human gesture segmentation based on change point model for efficient gesture interface. In: 2013 IEEE RO-MAN, South Corea, pp. 258–263 (2013)
14. Huang, D., Yao, S., Wang, Y., De La Torre, F.: Sequential max-margin event detectors. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014, Part III. LNCS, vol. 8691, pp. 410–424. Springer, Heidelberg (2014)
15. Chellali, R., Renna, I.: Emblematic gestures recognition. In: 2012 Proceedings of the ASME 11th Biennial Conference on Engineering Systems Design and Analysis (ESDA 2012). ASME ESDA 2012, vol. 2, pp. 755–753 (2012)