Robotics and Autonomous Systems 71 (2015) 134–149

Contents lists available at ScienceDirect

Robotics and Autonomous Systems

journal homepage: www.elsevier.com/locate/robot

Human motion based intent recognition using a deep dynamic neural model

Zhibin Yu, Minho Lee *

School of Electronics Engineering, Kyungpook National University, 1370 Sankyuk-Dong, Puk-Gu, Taegu 702-701, Republic of Korea

HIGHLIGHTS

- We developed an online deep dynamic neural model for intention classification.
- We evaluated the importance of internal action generation in motion classification and intention classification.
- Our proposed model performances better than a single layer supervised MTRNN.
- The possibility of each intention is able to be detected based on our model.

ARTICLE INFO

Article history:
Available online 19 January 2015

Keywords:
Human motion
Supervised MTRNN
Deep dynamic neural model
Motion classification
Intent recognition

ABSTRACT

The understanding of human intent based on human motions remains a highly relevant and challenging research topic. The relationship of the sequence of human motions may be a possible solution to recognize human intention. The supervised multiple timescale recurrent neural network (supervised MTRNN) model is a useful tool for motion classification. In this paper, we propose a new model to understand human intention based on human motions in real-time through a deep structure including two supervised MTRNN models, which are based on understanding the meaning of a series of human motions. The 1st supervised MTRNN layer classifies motion labels while the 2nd supervised MTRNN layer in the deep dynamic neural structure identifies human intention using the results of the 1st supervised MTRNN. We also considered the action–perception cycle effect between the 1st and the 2nd supervised MTRNNs, in which the motion label perception and internal action (motion prediction) form a cycle to improve the motion classification and intent recognition performance. A group of tasks was designed around movements involving two objects in an attempt to detect different motions and intentions based on the proposed deep dynamic neural model. The experimental results showed the deep supervised MTRNN to be more robust and to outperform the single layer supervised MTRNN model for detecting human intention. The action–perception cycle was found to efficiently improve both motion classification and prediction, which is important for human intent recognition.

1. Introduction

The recognition of human intent is a basic requirement for human–robot cooperation and interaction. If a robot is to understand and even predict human intention, it may be capable of providing assistance in the form of humanized services more promptly. For this purpose, various kinds of feature extraction methods are needed to find sufficient characteristics for analyzing and understanding human intention.

Generally, the vision-based feature extraction method has proven to be an efficient method for the interaction between human and machine [1]. Since advanced sensor systems such as Asus Xtion [2] offer us a convenient way to capture human motion, we are able to capture the human tester's skeletal information and record the position of each skeletal node in each frame. The sequential position of each skeletal node is used as initial input for our model.

Since human intention is not a momentary behavior but a continuous process, time series data are usually used for intention analysis [3,4]. The hidden Markov model (HMM) is considered to be an efficient dynamic tool to model and classify sequences of motions [5,6] and can also be used for intention recognition [7]. However, the HMM only considers the transitional probability of

* Corresponding author.
E-mail addresses: ahriman1985abr@gmail.com (Z. Yu), mholee@gmail.com (M. Lee).

http://dx.doi.org/10.1016/j.robot.2015.01.001
0921-8890/

each state. It cannot represent the contextual meaning of different motions and intentions. Further, it is difficult to measure the transitional probability between two kinds of motion that may be represented as variant time series data. The probability of two people performing the same motion would be different.

The recurrent neural network (RNN) model proposed by Husken and Stagge [8] showed another possible solution to dynamic signal prediction and classification. As there are two kinds of output neurons in this RNN model (prediction neurons and classification neurons), it is able to predict the output and classify the signals at the same time. Another RNN-based model developed by Yamashita and Tani [9], the multiple timescale recurrent neural network (MTRNN), proved to be efficient to predict and generate dynamic signals. The MTRNN model is developed based on a continuous timescale recurrent neural network (CTRNN) [10]. An interesting aspect of MTRNN is that this model is able to generate some untrained continuous signals based on existing knowledge [11,12]. It has been proven that the MTRNN is capable of efficiently predicting motion generation [13].

We considered the advantages of both models (MTRNN and RNN with two output layers) and developed a supervised MTRNN model that could be used for both motional classification and prediction in our previous work. Similar to the model of Husken and Stagge, both prediction and classification signals are generated simultaneously by the supervised MTRNN and can be used in a real-time process. The performance of supervised MTRNN in terms of motional classification has been proven to be efficient [14]. We wish to emphasize that this model has the ability to classify a lengthy untrained combination signal including several separately trained signals. Thus, this model is able to detect an unknown combination of motion signals if all elemental motion signals are trained. After we obtain the motion labels by analyzing the data sequence in each frame, we are able to get the intention labels by checking the data series between successive motions again. For this purpose, we need another supervised MTRNN model to detect the intention information based on the motion classification outputs.

An overview of our model is presented in Fig. 1. When a motion is observed, the model in the first MTRNN layer should recognize the performed motion. The motion label which is the output of the 1st layer will be reused as input for the 2nd layer and the intention label is obtained at the same time. Two different combinations of motional sequences may lead to two different intent recognition results even though some of their elemental motions are same. On the other hand, different intentions may end with the same motion. In this case, understanding the sequence of motions preceding the final motion is essential to recognize complex human intention.

In this paper, we considered eight kinds of meaningful motions and five kinds of human intentions. The motional classification ability, as well as the intention recognition performance is also evaluated. Moreover, the robustness of comparing our deep dynamic neural model with a single layer supervised MTRNN model to recognize human intention is also demonstrated.

Related work is introduced in Section 2. The proposed deep dynamic neural structure is introduced in Section 3. Section 4 presents the experimental results, which demonstrate that the proposed deep supervised MTRNN is able to classify different intentions as well as to distinguish between different human motions.

2. Related work

2.1. Encoding criteria for prediction and classification

We used Asus Xtion to extract skeletal nodes relating to human motion and to record their x and y position sequences. The normalization method was introduced in our previous work [14].

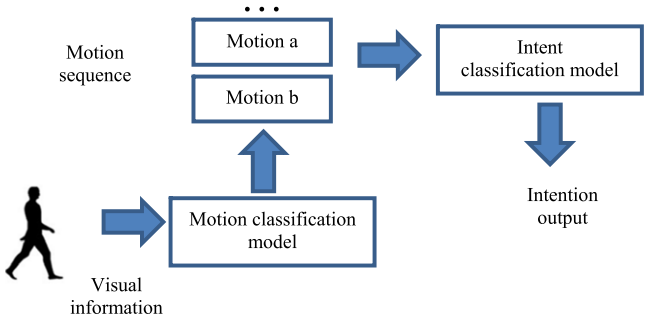

Fig. 1. Overview of motion-based human intent understanding.

The self-organizing map (SOM), which is commonly used as a preprocessing method for MTRNN feature extraction [11,12,15], is also used in our model. The input visual information is extracted using the following formula:

$$y_{i,t} = \frac{\exp\left(-\frac{\|v_{i,t}-v_{teach,t}\|^2}{\sigma}\right)}{\sum_{j\in V}\exp\left(-\frac{\|v_{j,t}-v_{teach,t}\|^2}{\sigma}\right)} \tag{1}$$

where v_i is the reference vector of ith node; $\|v_{i,t} - v_{teach,t}\|^2$ is the activation power of the ith node in time scale t; V means the whole SOM space; σ is a constant which is set to 0.01 and $y_{i,t}$ is an output vector which is used as the input for CTRNN.

The prediction output is calculated using the same SOM:

$$\hat{v}_{i,t+1} = \sum_{i\in V} y_{i,t} v_{i,t} \tag{2}$$

where $\hat{v}_{i,t+1}$ is the prediction output for the next step; $y_{i,t}$ is the activation output of the ith node of the CTRNN fast context layer.

There are two kinds of activation functions:

$$y_{i,t} = \begin{cases} \dfrac{\exp(u_{i,t})}{\sum_k \exp(u_{k,t})} & \text{if } i \in C, O \\ \dfrac{1}{1+\exp(-u_{i,t})} & \text{else} \end{cases} \tag{3}$$

$$u_{i,t} = \sum_j w_{ij} y_{j,t-1} \tag{4}$$

where $u_{i,t}$ is the input of ith node in time step t; $y_{j,t-1}$ is the state value of the jth node in time step $t-1$; w_{ij} is the weight from jth node to ith node; C and O represent the classification nodes and input–output nodes, respectively.

2.2. MTRNN

Context layers, which are the key components of MTRNN, are modeled by CTRNN. In comparison with RNN, CTRNN additionally considers the time scale effect. In a CTRNN, the output of each neuron is calculated using both the current input samples and the past history of the neural states. Hence, it makes the CTRNN suitable for predicting continuous sensori-motor sequences [16].

The MTRNN consists of two kinds of CTRNN: a "fast context layer" and a "slow context layer". The fast context layer, which has a smaller time constant, is capable of modeling elementary dynamic signals, and the slow context layer, which has a larger time constant, is believed to control the information sequence of the fast context layer and organize the overall signal sequence. The input–output layer obtains the information after SOM clustering and transmits it to the fast context layer. It can also pass the activation output from the fast context layer to the SOM layer