Neural Networks 92 (2017) 29–38

Contents lists available at ScienceDirect

Neural Networks

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

2017 Special Issue

Understanding human intention by connecting perception and action learning in artificial agents

Sangwook Kim[a], Zhibin Yu[b], Minho Lee[a,*]

[a] School of Electronics Engineering, Kyungpook National University, 1370 Sankyuk-Dong, Puk-Gu, Taegu 702-701, Republic of Korea
[b] College of Information Science and Engineering, Ocean University of China (OUC), 238 Songling Road Qingdao, China

ARTICLE INFO

Article history:
Available online 11 February 2017

Keywords:
Object-Augmented Multiple Timescale Recurrent Neural Network
Perception-action connected learning
Intention understanding
Human–robot interaction
Cognitive agent
Affordance

ABSTRACT

To develop an advanced human–robot interaction system, it is important to first understand how human beings learn to perceive, think, and act in an ever-changing world. In this paper, we propose an intention understanding system that uses an Object Augmented-Supervised Multiple Timescale Recurrent Neural Network (OA-SMTRNN) and demonstrate the effects of perception–action connected learning in an artificial agent, which is inspired by psychological and neurological phenomena in humans. We believe that action and perception are not isolated processes in human mental development, and argue that these psychological and neurological interactions can be replicated in a human–machine scenario. The proposed OA-SMTRNN consists of perception and action modules and their connection, which are constructed of supervised multiple timescale recurrent neural networks and the deep auto-encoder, respectively, and connects their perception and action for understanding human intention. Our experimental results show the effects of perception–action connected learning, and demonstrate that robots can understand human intention with OA-SMTRNN through perception–action connected learning.

1. Introduction

Intention understanding is important in developing artificial cognitive agents, such as robots, that can provide services to human beings (Blakemore & Decety, 2001). Recently, we have seen significant progress in the development of artificial agents (Kashiwabara, Osawa, Shinozawa, & Imai, 2012; Salamon, 2011), yet we are far from implementing some of the basic components of cognition, such as emotion and intention recognition. These mental abilities appear to be innate in human beings and render human–human interaction unique. In particular, the ability to understand the intention of others, has been argued to be the basis of human–human communication. Theory of mind (Premack & Woodruff, 1978) claims that human beings have a natural and inherent ability to understand the intention of others, or empathize with others. This ability, which allows humans to respond to each other in a coherent and useful manner, also extends to domains, that include language understanding, learning, and emotion recognition (El Ayadi, Kamel, & Karray, 2011). Therefore, to develop artificial agents that behave and respond similarly to intelligent humans, it is important to implement such mental abilities in them, especially the ability to understand the intention of others. However, to achieve this, we first need to explore the following: How do humans achieve this ability? What are the physiological and biological bases of empathy or the ability to understand the intention of others? Can we learn something from these cognitive processes? Moreover, can we simulate intention recognition ability in an artificial agent?

In this paper, we propose a system for understanding user intention and actively responding to provide suitable services. To implement intention recognition, we focus on two cognitive processes: the perception of object affordance and the prediction of human action. Both processes provide implicit and explicit information about intention, and they interact at various levels. Object affordance, a concept described by Gibson (1977), posits that objects are perceived based on the affordance of latent actions that can be performed with them. Action prediction, on the other hand, provides subtle clues as to the intention of the doer. Psychological and neurological studies have suggested that these two cognitive processes are not independent of each other; instead, they interact significantly at several levels in real-life scenarios. This interaction is often described as a perception–action connection and plays an important role in understanding the intention of others. We assume that

* Corresponding author.
E-mail address: mholee@gmail.com (M. Lee).

http://dx.doi.org/10.1016/j.neunet.2017.01.009
0893-6080/© 2017 Elsevier Ltd. All rights reserved.

implementing perception–action connections can be a key for implementing the ability to recognize human intention in an artificial agent.

At present, several computational models have been proposed for implementing intention recognition in an artificial agent. Some are based on object affordances, whereas others are based on action prediction. For instance, Kjellström et al. proposed an object categorization method based on the affordances between visual objects and simple human behaviors (Kjellström, Romero, & Kragić, 2011). In Montesano, Lopes, Bernardino, and Santos-Victor (2008), object affordances were modeled with Bayesian networks. These networks probabilistically represent dependencies that can be used to understand human actions. This understanding can be used to build robots that can imitate human behavior. Several stochastic models have also been adopted for intention recognition systems. The Hidden Markov Model (HMM) has been used to model the causality or dependency between successive measurements (Zhu, Cheng, & Sheng, 2008). Dynamic Bayesian Networks (DBN) have been used to model user intention (Schrempf & Hanebeck, 2005). DBN modeled the connections between intentions, and observed user actions and sensory modalities. The model obtained actions like explicit gestures, but did not consider objects related to actions.

Similarly, to extract intention cues from dynamic human actions, various dynamic models have been developed for managing action classification problems. HMMs are well known for analyzing and classifying human actions (Gehrig, Kuehne, Woerner, & Schultz, 2009), as are recurrent neural network (RNN)-based models (Hüsken & Stagge, 2003), such as Multiple Timescale Recurrent Neural Network (MTRNN) (Yamashita & Tani, 2008). This model was extended to a supervised MTRNN (SMTRNN) that was used for action classification (Yu & Lee, 2015).

The novel contribution of our system is that it attempts to integrate these two seemingly independent processes based on perception–action connections, into an Object Augmented-SMTRNN (OA-SMTRNN) for intention classification, assuming that these two processes interact at various levels of cognition and complement each other. We argue that determining intention based only on the affordance of perceived objects or predicted action sequences can be misleading because they can indicate several possible intentions. To correctly determine intention, an interaction between perception and action must be observed. We assume that perception and action together form a loop that improves the accuracy of intention recognition. To exploit the potential of OA-SMTRNN for intention classification, we conduct experiments using non-verbal intention classification tasks.

The remainder of this paper is organized as follows: in Section 2, we discuss the psychological and biological background and previous works. In Section 3, we describe OA-SMTRNN, the proposed computational model for understanding human intention. In Section 4, our experiments and their results are presented. Finally, we outline some concluding remarks and discuss future work in Section 5.

2. Psychological and neurological background and previous works

2.1. Intention

In psychology, intention refers to the idea or plan that an individual wants to implement. Theory of mind suggests that human beings have a natural and inherent ability to predict, represent, and interpret the intention of other human beings (Premack & Woodruff, 1978). This ability is also known as empathy. Humans recognize intention through various clues, such as language, body movement, facial recognition, gestures, and actions. Sometimes, a single clue is sufficient to indicate intention (usually in the case of language), but frequently, the combination of several clues confirms the intention. For example, consider the following scenario: a person approaches a table with multiple objects on it: a teapot, pen, paper, water jug, etc. It can be assumed that the person approaches with multiple intentions until she or he verbally specifies. However, without a verbal declaration, hand movements and the fact that this person is approaching the table could indicate that she or he has a particular intention. Perhaps this person wants to drink tea, take something off the table, drink water, or write. In this case, the person's intention can only be confirmed in relation to the object with which she or he interacts. If the person moves her or his hand to the teapot, it is possible that she or he wants to drink tea, offer it to someone, or take the teapot off the table. The action, in collaboration with the object with which the person interacts, restricts intentions and suggests possible actions. The person surely does not have the intention of writing because a teapot cannot be used to write. Moreover, once the person holds the teapot, the intention of whether she or he wants a drink, takes it off the table, or offers it to someone else is refined and tuned further by the person's action sequences. The most important point to note in this example is that the person's intention is inferred based on the cyclic interaction of actions and the perception of objects. Not only does action indicate intention, but objects themselves indicate possible actions that can be used to perform.

2.2. Perception–action connection

The classical approach to cognition has long assumed three stages of information processing: (1) perception, (2) cognition, and (3) action. In this model, perception and action do not interact directly, and additional cognitive processing is required to decide on an action based on perceptual information in a serial manner. This might require creating arbitrary linkages (mapping between sensory and motor codes). However, recent studies have shown that perception and action are not completely independent processes. Instead, they interact significantly at different levels. In the words of the famous psychologist James Gibson, "we must perceive in order to move, but we must also move in order to perceive". The idea that perception and action interact has been elaborated on in terms of a "perception–action cycle". Gibson (1950) described the concept of interactions between perception and action when he first indicated that, "when an observer moves relative to the environment, a global pattern of optical flow is generated at the moving point of observation and corresponds to the class and direction of observer movement. Reciprocally, this information can be used to regulate the forces applied by the observer in controlling subsequent movement, which in turn generates a new flow field and so on in a circularly causal connection, and it thus can rightly be called "movement-produced information". The idea of movement-produced information was later called a "perception–action cycle" and was further developed by Sperry (1969), who argued that the perception–action connection is the fundamental logic of the nervous system. Perception and action processes are functionally intertwined: perception is the means to action and action is the means to perception. Indeed, the vertebrate brain has evolved for governing motor activity with the basic function of transforming patterns into patterns of motor coordination.

2.3. Biological basis for a perception–action connection

With the development of brain imaging techniques, it has become possible to begin to understand the underlying neurological mechanisms of the interaction between perception and action.

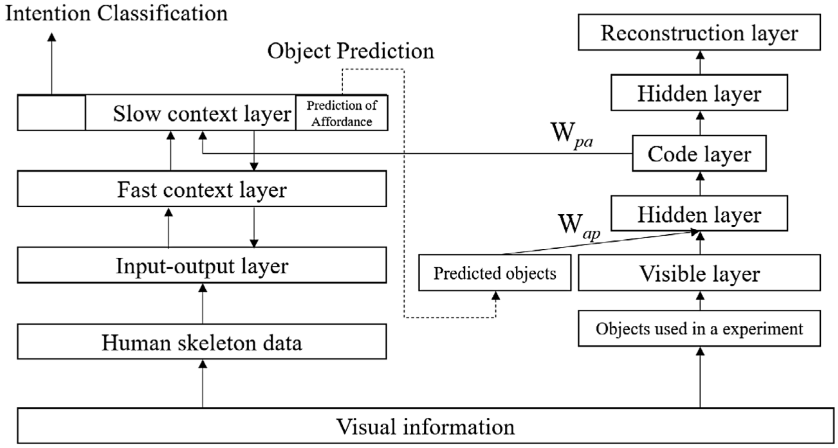


Fig. 1. Proposed OA-SMTRNN model for intention recognition.

Studies have shown that for all but the most primitive of animal species, these entire interactions are regulated by external and internal feedback (Fuster, 2003; Von Uexküll & Mackinnon, 1926). At all levels of the central nervous system, sensory-guided sequential action processing flows from posterior (sensory) to anterior (motor) structures, with feedback at every level. Thus, at cortical levels, information flows in a circular fashion through a series of hierarchically organized areas and connections that constitute the perception–action connection. Automatic and well-rehearsed actions in response to simple stimuli are integrated at low-level connections in posterior sensory areas (perceptual hierarchy) and motor areas of the frontal lobe (executive hierarchy). More complex behaviors, guided by more complex and temporally remote stimuli, require integration at higher cortical levels of both the perceptual and executive hierarchies, namely, higher sensory association areas and the prefrontal cortex (Fuster, 1988; Passingham, 1993).

Studies related to language perception have shown that understanding the meaning of words related to a motor action, such as "dance", requires far more than the well-known language areas of Broca and Wernicke in the left hemisphere of the brain. Hauk, Johnsrude, and Pulvermüller (2004) reported that the mere perception of action-related words also activates the motor homunculus—a cortical region of the brain that controls voluntary body movements. A later study by de Lafuente and Romo (2004) used fMRI to show that the processing of words related to an action performed with a body part, selectively activates the primary motor cortex according to the semantic content of the word, in addition to the conventional language areas. For instance, words such as lick, pick, and kick engage motor regions related to the control of the tongue, arm, and leg, respectively. These findings strongly suggest the interactive nature of perception (word or object) and action.

Support for the biological basis of perception–action connections also originates from the sensorimotor account of perception proposed by O'Regan and Noë (2001). Per this account, a component of meaning representation (of a word or object) incorporates sensorimotor contingencies – that is, information about the various ways in which the object can be acted upon and how its perceptual properties change in response to those actions – and various cognitive processes could refer to this representation as required. Similarly, the Common Coding Theory (Prinz, 1990) describes how perceptual and motor representations are linked; the theory claims that there is a shared representation (common code) for perception and action.

2.4. Previous works

Yamashita and Tani (2008) proposed the MTRNN brain model, and suggested that slow context nodes play a role similar to mirror neurons in the ventral premotor cortex (PMv) for action planning. They used different initial states for the slow context neurons as different goals to model a top-down process. In their research, the inferior parietal lobe (IPL) is modeled using fast context units and input–output units. The units in the input–output layers are used for both observation and prediction. The prediction output was used for robot behavior planning.

According to the hypothesis of MTRNNs, the slow context nodes are used to control the sequence of elemental actions saved in a fast context. Thus, the slow context nodes may contain different behaviors when different action signals are input. However, the slow context nodes of MTRNN are not designed for signal classification. In contrast with the work of Arie et al., which focuses on action generation only, supervised MTRNN for action understanding includes classification nodes in the slow context layer as well as action prediction nodes in the input–output layer (Yu & Lee, 2015). Classification nodes will continually fire following a bottom-up transmission if a known behavior is observed. The target value of classification nodes can be considered as a goal of prefrontal cortex (PFC).

The tree structure is often used to describe the composition of intent since different intentions are composed of different motional sequences. Our aim is therefore to determine human intent by observing human motional sequences and SMTRNNs are able to handle motion sequences by smooth compositionality which comes from their structure—multiple timescales. A key feature of a tree-like structure is that different leaves may have the same root, implying that different intentions may start with the same motion. Thus, knowledge of a human agent's current motion does not necessarily provide an understanding of the human agent's current intention and in those cases, information of object affordance can provide the important information for understanding of human intention.

3. Computational model for perception–action connected learning

3.1. Intention classification and object prediction based on understanding human action

We implement perception–action connected learning to understand human intentions using a newly proposed model, called OA-SMTRNN. The proposed model is an extension of our previous works (Kim, Kavuri, & Lee, 2013; Yu & Lee, 2015). The right side of Fig. 1 is modeled by a deep auto-encoder (Hinton & Salakhutdinov, 2006), which is used to analyze the relation information of objects.

The code layer in the deep auto-encoder is believed to refine information from visible layers, and represents latent human intention information correlated with objects selected by the user. The coded information is used as additional input for the slow context layer of the SMTRNN with weights W_{pa}. The SMTRNN, which is on the left side of Fig. 1, is still used to analyze the skeleton trajectories of human motion related to a specific intention. However, we would like the SMTRNN to classify human intentions as well as predict the selection of objects related to the current action. Thus, two different groups of classification nodes are defined for these two goals. With the help of the additional information from the code layer of the deep auto-encoder, the SMTRNN should be able to efficiently understand different intentions with similar actions but different objects. Additionally, the objects predicted by the SMTRNN can be reused as additional visible input to the hidden layer of the deep auto-encoder through the W_{ap} connection. Suppose that the numbers of neurons in a code layer of the deep auto-encoder is N_C; the first hidden layer of the auto-encoder contains N_H hidden neurons; the SMTRNN object prediction task includes N_O types of objects; and SMTRNN intention classification task has N_I intentions. Then, the size of W_{ap} is $N_O \times N_H$ and the size of W_{pa} is defined as $(N_S N_I N_O) \times N_C$, where N_S is the number of slow context nodes. There is no direct connection from the code layer to classification outputs so that code layer's outputs affect dynamics of the network indirectly rather than be used to classify intention or predict object directly. Otherwise, when code layer changes fast due to the noise or detection failures, it will affect intention classification and predictions of objects directly.

The arrows in Fig. 1 represent the flow of information. The arrows between the modules describe the path for perception–action connected learning.

Fig. 2 describes the execution flow of the proposed model in term of time index t. We call the left and right parts of the OA-SMTRNN the action and perception modules, respectively. When a human with an intention acts at time t, the left module of Fig. 1, the action module, tries to understand the meaning of human action and predicts an object label related to this intention. Then, the deep auto-encoder with the predicted object labels, the perception module, produces latent information in the code layer at time t; the code layer is connected to the slow context units and helps not only to identify human intention but also to predict the object selection at time $t + 1$.

We use the back propagation through time (BPTT) algorithm for training; the error function is defined using Kullback–Leibler divergence:

$$E = \sum_{t} \sum_{i \epsilon O} y^*_{i,t} \log \frac{y^*_{i,t}}{y_{i,t}}, \tag{1}$$

where O is a set of nodes in the input–output layer, $y^*_{i,t}$ is the desired output value of the ith neuron at time step t, and $y_{i,t}$ is the prediction value of the ith neuron with the existing weights and initial states.

The weight-updating rule is described in the following equation:

$$w_{ij}(n+1) = w_{ij}(n) - \alpha \frac{\partial E}{\partial w_{ij}}, \tag{2}$$

where n is the iteration step and α is the learning rate set at 0.0005 in our experiment. The partial differential $\frac{\partial E}{\partial w_{ij}}$ is given by

$$\frac{\partial E}{\partial w_{ij}} = \sum_{t} \frac{1}{\tau_i} \frac{\partial E}{\partial u_{i,t}} y_{j,t-1}. \tag{3}$$

The detailed structure of OA-SMTRNN action module is shown in Fig. 3. The concepts of the input–output and fast and slow context layers are inherited from the MTRNN model. According to Arie's work, the input–output layer of action module is modeled by a self-organizing map (SOM) to accept and output action sequences. The SOM algorithm preserves the topological properties of the input space. Although two SOM layers, including visual inputs and motor proprioception, are common in previous robot-related studies using MTRNNs (Yamashita & Tani, 2008), we did not use a real robot in our experiments. In our model, we combined the proprioception and vision (skeleton coordinates in the experiment) inputs into a single SOM layer. When visual information is obtained in time step t, SOM nodes are used for feature extraction. There is no connection between the input–output layer and slow context layer; the fast context layer works as a bond to connect them. The nodes within the fast and slow context layers are fully connected. The final outputs of the SOM nodes are used to calculate prediction error using BPTT.

Classification nodes in the OA-SMTRNN are the part of the slow context layer and have the same time constant, τ. The difference between the classification nodes and other slow context nodes is that the intention and object labels are used to train the classification nodes in the slow context unit. Classification nodes need to back propagate not only an action prediction error, but also an intention inference and object prediction error to other nodes. All nodes, including the classification nodes, work synchronously. When a test sequence is given, the classification nodes that correspond to the label are activated.

Considering the classification error, the partial differential equations are defined as:

$$\frac{\partial E}{\partial u_{i,t}} = \begin{cases} y_{i,t} - y^*_{i,t} + \left(1 - \frac{1}{\tau_i}\right) \frac{\partial E}{\partial u_{i,t+1}} \quad i \in O \\ \left(y_{i,t} - y^*_{i,t}\right) f'\left(u_{i,t}\right) + \sum_{k \in C} \frac{\partial E}{\partial u_{k,t+1}} \\ \qquad \times \left[\delta_{ik}\left(1 - \frac{1}{\tau_i}\right) + \frac{1}{\tau_k} w_{ki} f'\left(u_{i,t}\right)\right] \\ \quad i \in C_{object}, C_{intention} \\ \sum_{k \in N} \frac{\partial E}{\partial u_{k,t+1}} \left[\delta_{ik}\left(1 - \frac{1}{\tau_i}\right) + \frac{1}{\tau_k} w_{ki} f'(u_{i,t})\right] \\ \quad \text{otherwise} \end{cases}, \tag{4}$$

where $f'(x)$ is the derivative of the sigmoid function, $y_{i,t} - y^*_{i,t}$ is the difference between the neuron output and the ideal value, $u_{i,t}$ is the ith neuron state at time step t, τ_i is the constant that controls the neuron state update speed, and δ_{ik} is the Kronecker delta ($\delta_{ik} = 1$ if $i = k$; otherwise, 0). O indicates a set of input–output nodes, and C_{object} and $C_{intention}$ represent the classification nodes used for object prediction and intention classification, respectively, in the slow context layer. The action prediction signal is generated in the input–output layer while the classification task is performed in the slow context layer. However, we modified this equation for object prediction. Instead of using the softmax activation function, we use the sigmoid function to calculate the final output, so that multiple objects can be used for one intention. It must be noted that the softmax function only supports the classification which gives one category as an output.

Using prediction result from the OA-SMTRNN action module, we can improve the auto-encoder with the following equation:

$$y_h = f\left(\sum_{V} w_{vh} v + \sum_{A} w_{ap} y_a + b\right) \tag{5}$$

where y_h is the hidden neuron output, w_{vh} is the connection between the visible and current hidden layer, w_{ap} is the connection from the action module to the perception module, y_a is the postsynaptic value of the object prediction outputs based on human action understanding and b is a bias.

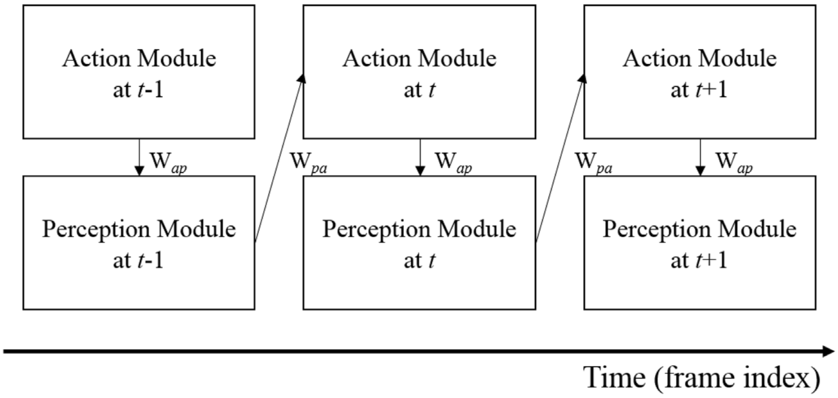


Fig. 2. A brief execution flow of the proposed model.

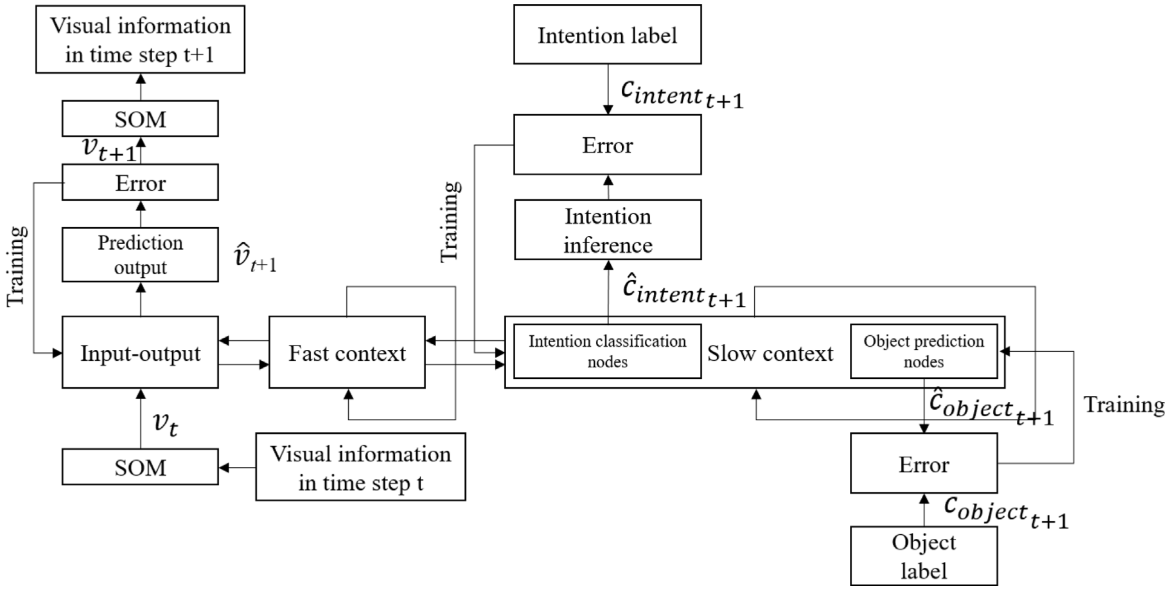


Fig. 3. Action module for intention inference and object prediction based on action understanding.

The training of w_{ap} can still follow the error back-propagation rule:

$$\frac{\partial E}{\partial w_{ap}} = \frac{\partial E}{\partial u_h} y_a \tag{6}$$

where u_h is the presynaptic value of a neuron in the auto-encoder.

The predicted object labels are contextual biases for the perception module implemented by the deep auto-encoder. Subsequently, the new input vector to the perception module becomes

$$\mathbf{v}_{appended} = \{\mathbf{v}_{original}, \mathbf{y}_a\}, \tag{7}$$

where $\{,\}$ indicates vector concatenation, $\mathbf{v}_{original}$ is the original visible nodes of the deep auto-encoder, and $\mathbf{y}_a$ is the object prediction outputs from the OA-SMTRNN action module.

Whereas nodes for intention inference and object prediction are located in the slow context layer, action prediction signals are generated in the input–output layer, which has a small time constant because it should be as fast as the original action sequences.

As we mentioned earlier, we place the prediction of object labels from the OA-SMTRNN action module as additional visible nodes in the bottom layer of the deep auto-encoder. Consequently, the number of weights from the visible to first hidden layer increases. Therefore, when we obtain a new sequence, the OA-SMTRNN action module is used to predict possible objects based on the current and past action sequences. We assume that the additional information can help the perception module exclude irrelevant objects and reconstruct object labels in cases of confusing or incorrect object detection.

3.2. *Affordance modeling*

As previously mentioned, user intention can also be inferred based on affordance. In the proposed OA-SMTRNN, affordance is modeled using a deep auto-encoder (Kim et al., 2013), because deep-structured networks are known for their ability to capture high-order relationships or features that are unobservable directly. Auto-encoder models usually consist of two parts: the encoder and decoder. An encoder transforms the original input signal into relatively short codes, which can be viewed as compression. A decoder is the counterpart of the encoder and reconstructs the original information from the encoded signal. In previous work that refers to this model, an auto-encoder input is a vector that represents affordance information. The affordance information is encoded, and the related objects are decoded (see Figs. 4 and 5).

In general, the task of an auto-encoder can be viewed as meaningless because it unnecessarily compresses information and recovers it again. However, auto-encoders can extract useful information from signals and remove noise if the auto-encoder

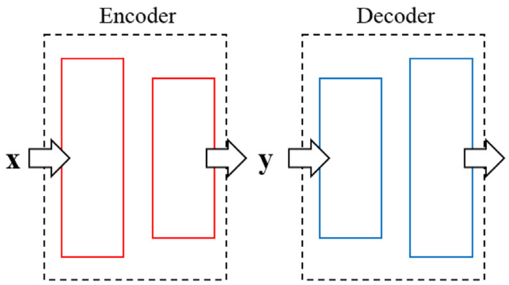


Fig. 4. Diagram of auto-encoders.

is appropriately constructed. In our case, we need the auto-encoder to analyze object affordance regarding human intention and extract meaningful latent information from the observed objects to help with intention inference.

Because of their deep structure, training auto-encoders is difficult. In our model, we use the method proposed by Hinton and Salakhutdinov (2006), building Restricted Boltzmann Machines (RBM) in a layer-by-layer manner to initialize the encoder network. A decoder network is simply built by reversing the encoder and fine-tuning it.

Error back-propagation, a dominant training method for artificial neural networks with differentiable activation functions, is used in the fine-tuning procedure for the auto-encoder network.

After the auto-encoder network has been built, the output vector produced in a code layer of this perception module contains sufficient object affordance information for reconstruction. When a person has a certain intention, she or he often employs a corresponding object. The latent code vector can be used to incorporate information about the co-occurrence of objects. Therefore, the code, information of object affordance, inhibits irrelevant intentions and indicates relevant intentions as more likely.

At the uppermost layer, the objects related to the affordance code are reconstructed. Even if the information on an object is missing from the original vector because of noise, it can be revealed through decoding the affordance code. This characteristic can yield a prediction for the future state of objects, even if the deep auto-encoder model is static, because object selection occurs serially, and missing information at the current time can be predicted by affordance and its decoding.

Although the model reconstructs related objects based on visual information, sometimes noise prevents visible nodes from obtaining correct information. In this case, predicted object labels from the OA-SMTRNN action module in Fig. 6 can help make the model robust to noise. The additional nodes predicting object labels are connected to the hidden layer as the dynamic biases and weights between predicted nodes and the hidden layer are trained. At the same time, the other weights are fine-tuned with the additional nodes to reconstruct proper information. Though the decoder is required for training the deep auto-encoder, the code layer's outputs in the perception module (the deep auto-encoder), are considered to be an additional input to the slow context layer of the action module. The weights between the code and slow context layers are fully connected and are updated with Eq. (3) using back-propagation with the Kullback–Leibler divergence error from Eq. (1).

Information from the affordance model to the action module, represented as the arrow from the code layer to the slow context layer in Fig. 1, enriches the features required for accurately recognizing human intentions. Originally, the MTRNN and SMTRNN predicted action sequences based on the dynamics of latent states through time. Here, however, by integrating such information, the intention classification functionality of the OA-SMTRNN can consider skeletal dynamics and object affordance simultaneously in a unified framework.

Because the perception module of object affordance is connected to the slow context layer, the information affects the input–output layer which predicts the action sequences. Therefore, perception information helps the module suppress irrelevant actions. If several intentions share similar dynamic characteristics, lower layers of the action prediction module provide similar outputs, i.e., their features are hardly distinguishable; however, the encoded feature information from the affordance perception module makes the data easily separable by inserting feature dimensions related to object affordances. We expect that the information saved in the code layer helps OA-SMTRNN predict action sequences accurately and improve its intention inference performance.

4. Experiment results

4.1. Simulation of intention understanding

To test the performance of our proposed model, we conducted an experiment with a dataset based on six intentions and eight objects. A total of 6 × 16 (skeleton + image) sequences were captured using a Microsoft Kinect 2 at approximately 20–30 fps. A total of 12 skeleton points were used in our experiment (Spine Mid, Neck, Head, Shoulder Left, Elbow Left, Wrist Left, Shoulder Right, Elbow Right, Wrist Right, Hip Left, Hip Right, and Spine Shoulder). Thus, the input dimension for the action module is 24 (x- and y-axes). A total of 6 × 12 sequences were used for training, and the rest were used for testing. Each sample lasted approximately 150–300 frames. The experiment is further described in Table 1.

As mentioned earlier, multiple intentions can be inferred from the same action. It is also possible for objects employed by a user to be difficult to identify because of occlusion (Fig. 7). This situation poses the problem of intention inference by only considering object information, which we aim to solve using the advantages of the OA-SMTRNN. We expect that the output of the

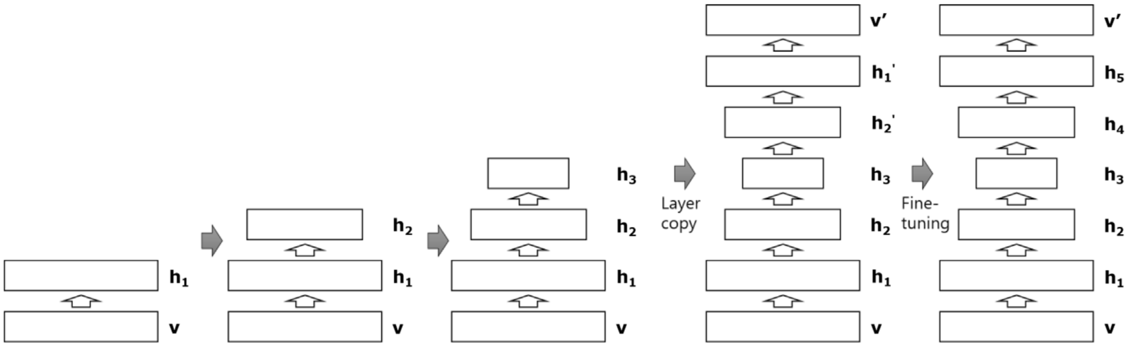


Fig. 5. Training procedure for deep auto-encoder.

Table 1
Intention description.

Intention	Objects included	Actions included
Drinking espresso	Espresso	Grasp one object, take something to mouth
Drinking milk	Milk carton	Grasp one object, take something to mouth
Reading a book	Book	Grasp one object, hold and read
Eating noodles	Instant noodles	Grasp one object, hold and eat
Making an americano	Cup, espresso, water	Grasp two objects, pour something from one to another, put down one object, grasp another object and pour something into the other.
Making a café mocha	Cup, espresso, chocolate powder	Grasp two objects, pour something from one to another, put down one object, grasp another object and pour something into the other.
Making a cup of tea	Cup, water, tea	Grasp two objects, pour something from one to another, put down one object, grasp another object and pour something into the other.
Making a cup of water	Cup, water	Grasp two objects, pour something from one to another

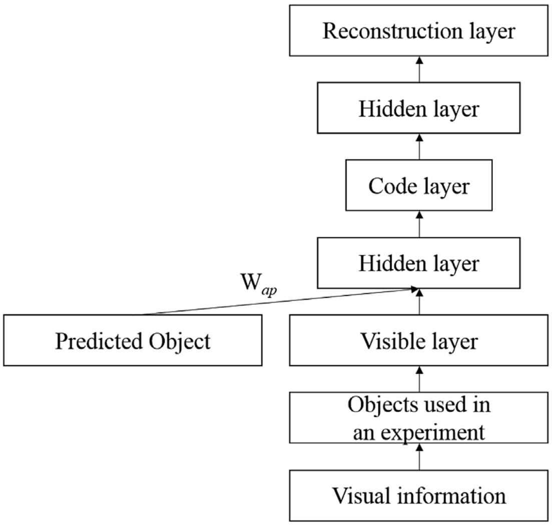


Fig. 6. Structure of OA-SMTRNN perception module.

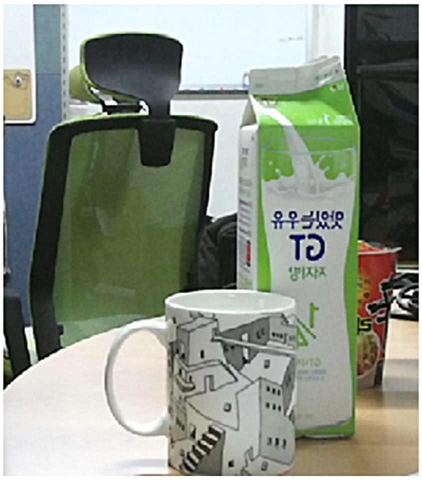


Fig. 7. A case of overlapping objects.

OA-SMTRNN action module can help the perception module to reconstruct correct candidate objects labels. To recognize objects that are grasped by the hand, we used the SURF detector (Bay, Ess, Tuytelaars, & Van Gool, 2008) based on a saliency map (Jeong, Ban, & Lee, 2008) around the hand coordinates from the Kinect 2. BPTT (for the action module) and Contrastive Divergence + BP (for the perception module) were used to train the modules separately until they converged. Finally, the modules were fine-tuned with the connection shown in Fig. 1.

We used the auto-encoder setting from our previous work (Kim et al., 2013). However, the objects we used in this study overlapped with each other, as shown in Fig. 7. In such cases, the object information that feeds into the auto-encoder might not be correct. Given incorrect or confusing information, the auto-encoder cannot achieve good results. However, the OA-SMTRNN action module, which only considers action information, can solve this problem by considering the relationship between the action and objects. Fig. 8(a) and (b) show two examples obtained with a "making a cup of water" sequence. Fig. 8(a) shows the intention inference outputs over time in the slow context layer. The intention output for making a cup of water starts to grow after approximately the 20th frame. Fig. 8(b) shows the objects label prediction outputs. The output neurons that correspond to the cup and water fire in sequence. Fig. 8(c) and (d) show a case that is more complex than the previous one. The user grasps the cup and chocolate first. After pouring the chocolate powder into the cup, the user replaces the chocolate powder and grasps the coffee container to make a café mocha. According to this action sequence, the neurons that correspond to making a cup of espresso fire at an early stage because at this moment, the user performs steps similar to making an americano. Only after the system observes another pouring action, do the neurons that correspond to making a café mocha fire and confirm the intention (Fig. 8(c)). Fig. 8(d) shows the object prediction outputs. The neurons that correspond to three types of objects fire sequentially, meaning that these three objects should be used for the current intention (making a café mocha).

As a supplement, we also conducted an experiment to show whether object affordance information helps with action-based intention classification, the results of which are listed in Table 2. In the very early stages of the action sequence, the person has not yet moved; without any information, the model cannot infer intention. Therefore, the accuracies are calculated by the averages over frames after the first 70 frames. Without object affordance information, the system might not be able to understand exactly what the tester is attempting to do. Since they have very similar actions, they are not able to be distinguished and one action goes totally unrecognized. Even for the connection model, "making a cup of tea" and "making a café mocha" are very similar until the person picks an object such as tea or coffee. In some cases, action sequences are enough to distinguish the intention correctly. For example, "eating noodles" has a unique eating motion that can be recognized accurately without any object affordance information. Fig. 9 shows the confusion matrix for intention classification results of the SMTRNN and OA-SMTRNN. In Fig. 9(a), for the SMTRNN, "making a cup of tea" is confused heavily with "making a café mocha" and "drinking espresso" fails to be distinguished from other intentions —"drinking milk", "reading a book", or "making a cup of water".

Fig. 10 shows the entropy change curves for 12 different intention sequence samples. In most cases, the red curves that

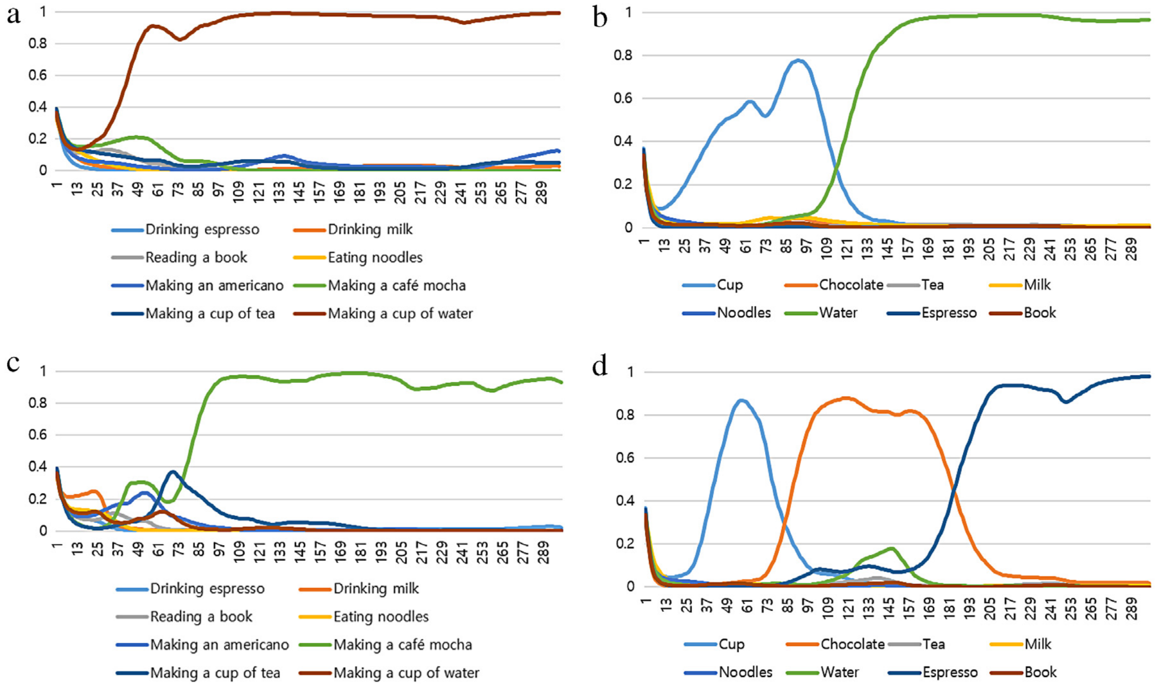
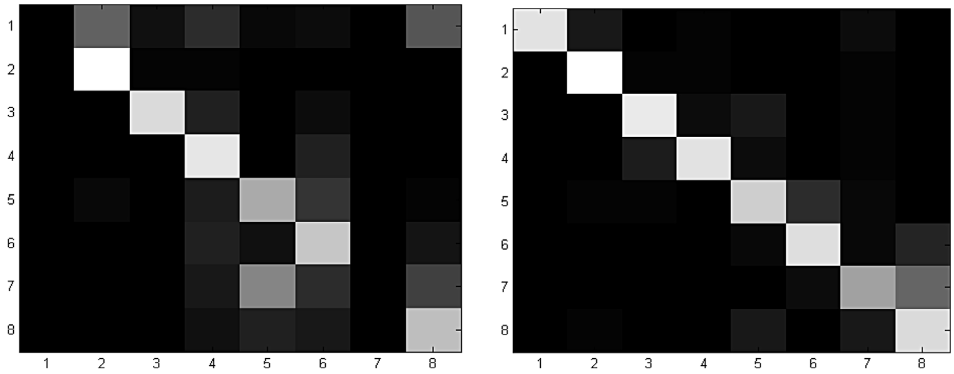


Fig. 8. Real-time output of the OA-SMTRNN. (a) Sample action-based intention outputs when making a cup of water; (b) Object label predictions using (a); (c) Sample action-based intention outputs when making a café mocha; (d) Object label prediction using (c).



(a) SMTRNN.

(b) OA-SMTRNN.

Fig. 9. Confusion matrix for intention classification results.

Table 2
Real-time intention classification results.

Intention	SMTRNN (true positive %)	OA-SMTRNN (true positive %)
Drinking espresso	0	81.83
Drinking milk	94.67	92.00
Reading a book	80.67	83.83
Eating instant noodles	85.50	81.83
Making an americano	62.67	73.42
Making a café mocha	72.75	79.83
Making a cup of tea	0	58.50
Making a cup of water	70.50	78.42
Average	58.35	78.71

show the cases with the connected model drop faster and reach robust states. However, without the connection, entropy is reduced slowly or cannot reach a stable state (cases (k) and (l)).

4.2. Experiments in a café environment

We acquired ordering situations from several café chain stores in our university for the experiments with real café data. The corpus contains cases of interactions with different intentions. In most cases, there is a customer and a clerk who take turns in the interaction. For simplicity, we exclude a few cases in which multiple customers order together. Because the model aims to recognize the customer's intentions on behalf of the clerk, the clerk's behaviors are removed in every case. The sequences of skeleton points are obtained using Kinect v2.

In this case, the interaction sequence is important context information. To consider this in the model, we split the corpus

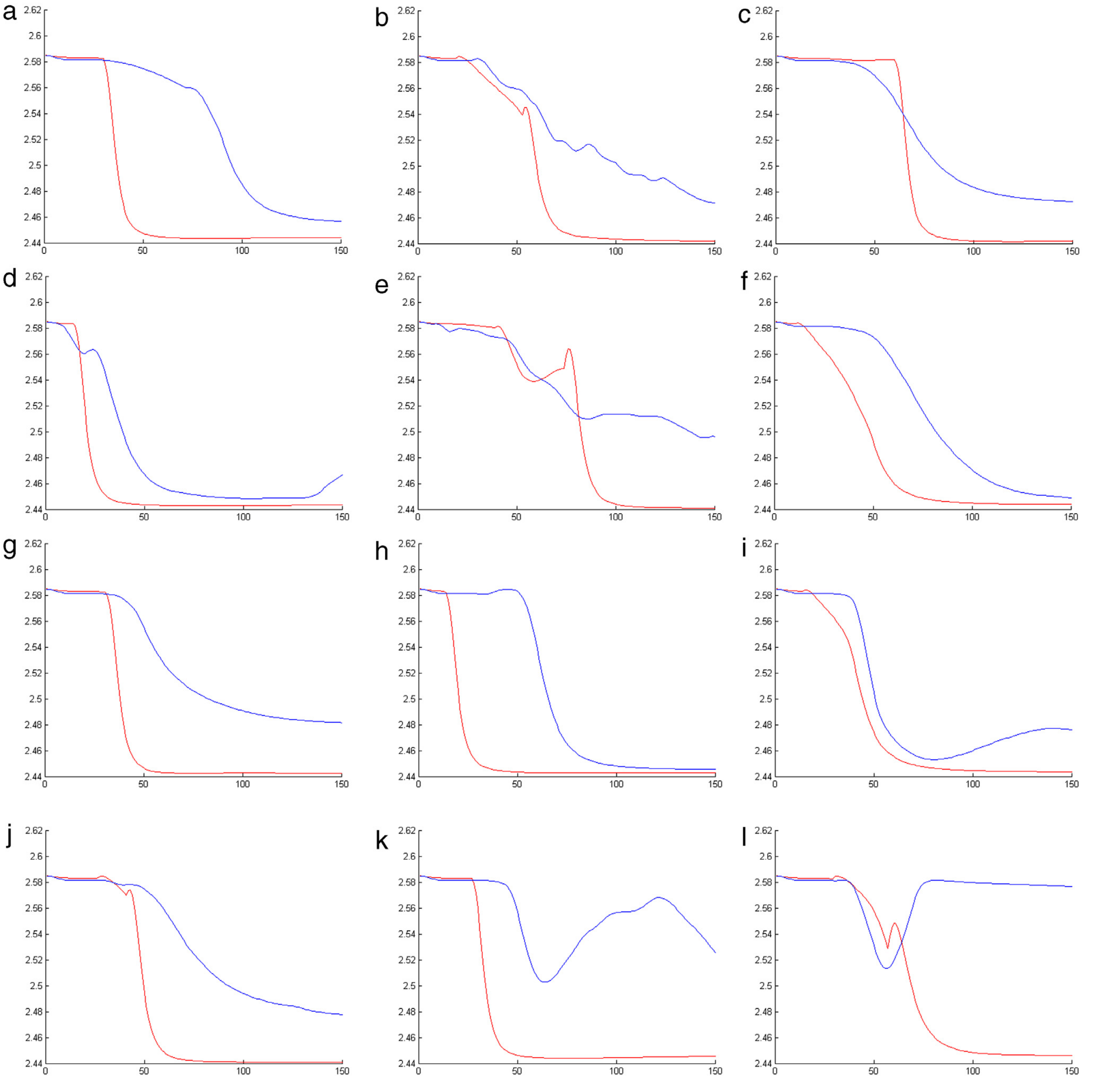


Fig. 10. Real-time entropy change curves for 12 intention sequence samples. Blue: entropy reduction without additional object affordance; Red: entropy reduction with additional object affordance. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

dataset into 5 sequence sets. The number of intention in the 5 sequence sets is 10, 14, 13, 7 and 8. A total of 57 × 3 sequences were captured. 57 × 2 sequences were used for training, and the remaining set was used for testing. Each sample lasted approximately 150–300 frames. Each set consists of sequences of user actions captured from the Kinect and objects corresponding to the action, if applicable. For example, sequence set 1 includes the first interaction in every order, such as a greeting, and sequence set 2 has sequences from the second turn of the interactions, such as simple ordering for some drinks. Sequence set 3 includes additional requirements by guests such as coffee size and sugar preferences. Sequence set 4 is usually related to payment. Sequence set 5 is related to the end of ordering and farewells. Each OA-SMTRNN tries to understand the human intention for each sequence, and five OA-SMTRNNs cascading through all of the sequences determine the guest's consecutive intentions in the café. Fig. 11 describes this configuration.

Table 3
Real-time intention classification results in a café environment.

Intention	SMTRNN (true positive %)		OA-SMTRNN (true positive %)	
	Training set	Test set	Training set	Test set
Sequence 1	90.57	80.1	98.43	97.87
Sequence 2	6.97	6.86	98.17	90.00
Sequence 3	7.77	6.62	98.08	96.87
Sequence 4	97.76	28.38	94.33	91.81
Sequence 5	21.03	19.25	95.21	92.33
Average	44.82	28.24	96.84	93.78

Table 3 shows the accuracies of each interaction step. For all test sets, the OA-SMTRNN shows significantly better result than

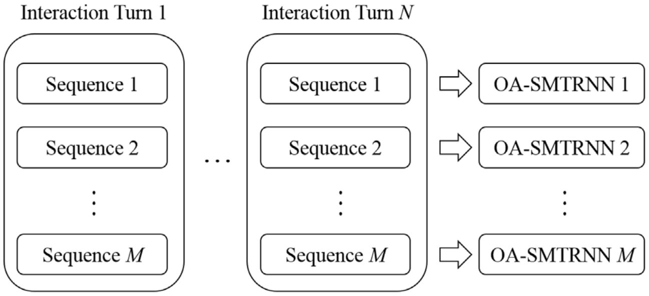


Fig. 11. Auto-encoder constructions considering the context in sentence sequences.

the SMTRNN-only model. As shown in the table, sequences 2, 3, and 5 have failed to converge properly without the proposed connections. For sequence 4, the SMTRNN shows better accuracy (97.8%) than the OA-SMTRNN over the training set, but it cannot generalize well over the test dataset.

5. Concluding remarks

In this paper, we proposed the development of an artificial agent that can infer human intention based on perception–action connected learning. The new dynamic neural model for intention classification used in this paper is an object-augmented SMTRNN that consists of two different artificial neural network models that incorporate action and perception information and connects them to understand human intentions more precisely. The OA-STMTRNN can manage action and object prediction, as well as intention inference. The proposed model offers a possible method for considering affordance based object information and action prediction interactively to predict human intention accurately in human–robot interaction. The experimental results show that action and object prediction in the action module can help the object affordance module find latent information regarding human intention. The code layer information in the perception module provides additional conditions to the action module, which is critical for classifying human intention. Based on our results, we argue that implementing perception–action connected learning can play an important role in developing artificial agents that can infer human intention and interact better.

However, there is still significant work to complete before releasing this system for real-world applications. Object detection and recognition are still challenging to perform in real time. Convolutional neural networks, which are good at recognizing object shapes, could be a solution to this problem. Moreover, lower-level perception–action connections that embrace the object recognition modules like those of convolutional neural networks can also be considered. Currently, the proposed model has only perception–action connections; however, we will consider a cyclic perception–action connection that allows recursive interaction between perception and action modules, which may improve human intention classification performance. In addition, the use of large-scale corpus acquired from an external source (e.g., Wikipedia) to pretrain the affordance model can be considered in future works.

Acknowledgments

This work was supported by the Industrial Strategic Technology Development Program (10044009) funded by the Ministry of Trade, Industry and Energy (MOTIE, Korea).

References

Bay, H., Ess, A., Tuytelaars, T., & Van Gool, L. (2008). Speeded-up robust features (SURF). *Computer Vision and Image Understanding*, *110*, 346–359.

Blakemore, S.-J., & Decety, J. (2001). From the perception of action to the understanding of intention. *Nature Reviews Neuroscience*, *2*, 561–567.

de Lafuente, V., & Romo, R. (2004). Language abilities of motor cortex. *Neuron*, *41*, 178–180.

El Ayadi, M., Kamel, M. S., & Karray, F. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, *44*, 572–587.

Fuster, J. M. (1988). Prefrontal cortex. In *Comparative neuroscience and neurobiology* (pp. 107–109). Springer.

Fuster, J. M. (2003). *Cortex and mind: Unifying cognition*. : Oxford university press.

Gehrig, D., Kuehne, H., Woerner, A., & Schultz, T. (2009). Hmm-based human motion recognition with optical flow data. In *2009 9th IEEE-RAS international conference on humanoid robots* (pp. 425–430). IEEE.

Gibson, J.J. (1950). The perception of the visual world.

Gibson, J. (1977). *The theory of affordances. perceiving, acting, and knowing. and knowing*. Hillsdale: Lawrence Erlbaum Associates.

Hauk, O., Johnsrude, I., & Pulvermüller, F. (2004). Somatotopic representation of action words in human motor and premotor cortex. *Neuron*, *41*, 301–307.

Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science*, *313*, 504–507.

Hüsken, M., & Stagge, P. (2003). Recurrent neural networks for time series classification. *Neurocomputing*, *50*, 223–235.

Jeong, S., Ban, S.-W., & Lee, M. (2008). Stereo saliency map considering affective factors and selective motion analysis in a dynamic environment. *Neural Networks*, *21*, 1420–1430.

Kashiwabara, T., Osawa, H., Shinozawa, K., & Imai, M. (2012). TEROOS: a wearable avatar to enhance joint activities. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 2001–2004). ACM.

Kim, S., Kavuri, S., & Lee, M. (2013). Intention recognition and object recommendation system using deep auto-encoder based affordance model. In *The 1st international conference on human-agent interaction*.

Kjellström, H., Romero, J., & Kragić, D. (2011). Visual object-action recognition: Inferring object affordances from human demonstration. *Computer Vision and Image Understanding*, *115*, 81–90.

Montesano, L., Lopes, M., Bernardino, A., & Santos-Victor, J. (2008). Learning object affordances: From sensory–motor coordination to imitation. *IEEE Transactions on Robotics*, *24*, 15–26.

O'Regan, J. K., & Noë, A. (2001). Acting out our sensory experience. *Behavioral and Brain Sciences*, *24*, 1011–1021.

Passingham, R. E. (1993). *The frontal lobes and voluntary action*. Oxford University Press.

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind ? *Behavioral and Brain Sciences*, *1*, 515–526.

Prinz, W. (1990). A common coding approach to perception and action. In *Relationships between perception and action* (pp. 167–201). Springer.

Salamon, T. (2011). *Design of agent-based models*. Eva & Tomas Bruckner Publishing.

Schrempf, O.C., & Hanebeck, U.D. (2005). A generic model for estimating user intentions in human–robot cooperation. In *ICINCO* (pp. 251–256).

Sperry, R. W. (1969). A modified concept of consciousness. *Psychological Review*, *76*, 532.

Von Uexküll, J., & Mackinnon, D. L. (1926). In K. Paul (Ed.), *Theoretical biology*. Trench: Trubner & Company Limited.

Yamashita, Y., & Tani, J. (2008). Emergence of functional hierarchy in a multiple timescale neural network model: a humanoid robot experiment. *PLoS Computational Biology*, *4*, e1000220.

Yu, Z., & Lee, M. (2015). Real-time human action classification using a dynamic neural model. *Neural Networks*, *69*, 29–43.

Zhu, C., Cheng, Q., & Sheng, W. (2008). Human intention recognition in smart assisted living systems using a hierarchical hidden markov model. In *2008 IEEE international conference on automation science and engineering* (pp. 253–258). IEEE.