Intention Recognition and Object Recommendation System using Deep Auto-encoder Based Affordance Model

Sangwook Kim[1], Swathi Kavuri[2], and Minho Lee[1]

[1]School of Electronics Engineering, Kyungpook National University,
[2]The Institute of Electronic Techonology, Kyungpook National University,
1370 Sankyuk-Dong, Puk-Gu, Taegu 702-701, Republic of Korea

Abstract: Intention recognition is an important task for human-agent interactions (HAI) since it can make the robot respond adequately to the human's intention. For the robot to understand the world in terms of its own actions, the robot requires the definition of adequate knowledge representations. Affordance is the concept used to represent the relation between an agent and its environment. A robot can exploit this type of knowledge to infer implicit human intentions. In this paper, we propose a system based on action-object affordances modeled using deep structure that can recognize the user's intention and recommend the corresponding objects related to that intention. The network is learnt by the robot after considering the user's attention for specific objects. To notice the user's attention, the gaze information is obtained using Tobii 1750 eye-tracker in experiments. The experimental results show the successful recognition and recommendation performance of the proposed system.

1 Introduction

In cognitive psychology, intention refers to an idea or plan of what we are going to do. According to theory of mind [1], human beings have a natural way to predict, represent and interpret user intention reflected implicitly or explicitly by the others. Thus, humans can serve the customer by understanding their attention for specific objects. But for the robot to act in a complex world and understand user's intention, the robot requires the definition of adequate knowledge representations to support the execution of a large number of tasks.

Affordance is the concept defined by Gibson [2] which represents the relationships or possibilities between actions and objects. An affordance is an intrinsic property of an object, allowing an action to be performed with the object. A robot can exploit this type of knowledge to understand the behavior of the world in terms of its own actions and can infer the user's intention with objects more easily. In [3], authors proposed object categorization method based on affordances between visual objects and actions obtained from human demonstration. And in [4], object affordances were modeled with Bayesian networks which are the probabilistic representation of dependencies and used to understand actions and imitate the human's behavior by a robot.

Also, various stochastic models have been adopted for intention recognition system. Hidden Markov model (HMM) was used as the recognition model [5] to model the causality or dependency between successive measurements. Dynamic Bayesian networks (DBN) was also used to model user's intention [6], it was adopted in a hybrid form which treats continuous and discrete valued states in a model. It modeled connections among intentions, observed user actions and sensor modalities. They obtained actions like explicit gestures but didn't consider about neither the user's attention nor objects related to actions.

In [7], authors briefly introduced their works to recognize human intention by analyzing the change in distance between the observed human's hand and the objects in the scene over several frames. By using stacked denoising auto encoder (SDA) which learns distances of objects and predicted object for which the person is currently reaching. But the system just tried to understand current user's behavior and couldn't recommend the objects needed for the user.

In [8], authors classified intentions given objects using naïve Bayes classifier [9] and used eye-tracking data as an input. But their model only classifies intentions, so couldn't explain affordances between intentions and objects, and recommendation of objects was just assumed to be obtained indirectly by sending a query

containing the intention to imaginary database. The application we are focusing on is robot's learning of objects related to user's specific intention to perform a particular action. For example, from a kitchen scene containing different objects such as ramen box, an instant coffee box, knives, an electric kettle, a pot, a gas stove, beer cans, mineral waters, glasses, breads, wines, and mugs, the proposed system recommends objects related to specific intensions such as 'eat noodle', 'drink beer', 'drink water' and 'eat bread' etc.

In this paper, we propose the system that recognizes the human intention and recommends corresponding objects based on objects of attention and intention-object affordances modeled by using a deep auto-encoder. Since stacked denoising auto-encoder could perform the robust classification and reconstruction with denoising , it is adequate to recognize the intention and to recommend related objects of high affordance corresponding to the intention. Affordances appear from the interaction between the robot and the environment. In order to model, acquire and use affordances, we obtain the attention of human first. One can usually get this information about the environment from his/her eyes, since gaze information of the user is crucial cue of his/her attention. This kind of joint attention is very important to recognize human intention [10]. For example, if the user stares a certain object longer than his ordinary gazing time, it can be interpreted as his interest or attention on that object [11, 12]. In [13, 14], intentions of the user are categorized to navigational intent and informational one and those intentions are recognized by obtaining information about user's eye including pupillary response and gaze information without considering object which he/she wants.

The rest of this paper is organized as follows. In Section 2, we describe the overall structure of the proposed model and discuss the deep auto encoder used. In Section 3 we present the experimental results to evaluate the performance of the proposed. Finally, we draw our conclusions in Section 4.

2 Proposed Model

2.1 Overall Structure

The proposed system has two major functions, intention recognition and related object recommendation. In this paper, we use the deep auto-encoder to encode the intention and model affordances based on weights. Overall structure is shown in Fig. 1. It is the encoder that predicts the user's intention based on Markov property. In other words, if the intention depends on objects viewed for a fixed period, the problem of intention recognition can be defined as:

$$\hat{I} = argmax_I \, P(I|\{obj_{attention}\}), \quad (1)$$

where I is the pre-existing information on intention, $\hat{I}$ is the predicted intention, $\{obj_{attention}\}$ means the set of objects of attention and $P(\cdot)$ is the function which evaluates the possibility of the intention.

From the predicted intention, the decoder calculates the action-object affordances given all of objects observed in a scene. The system then compares the decoder output with the objects of attention to recommend objects to the user based on Eq. (2).

$$\{obj_{recommended}\} = f\left(A(\hat{I}), \{obj_{attention}\}\right), \quad (2)$$

where $A(\cdot)$ denotes the affordance function between intentions and objects, which yields the intention-related objects as the result, and $f(\cdot)$ evaluates labels of the recommended objects $\{obj_{recommended}\}$, based on the $A(\hat{I})$ and objects of attention $\{obj_{attention}\}$. Through the deep auto-encoder, user's intention is predicted by the encoder and the decoder part reconstructs objects which have affordances with the predicted intention.

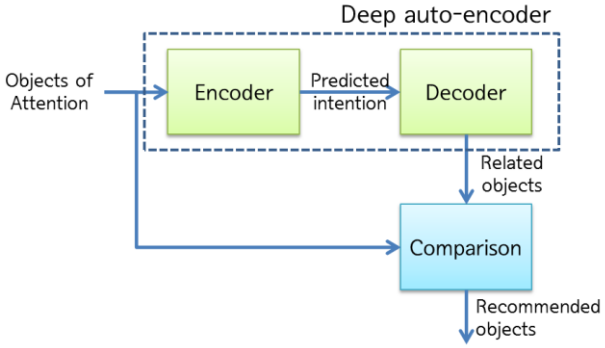

Fig. 1. Overall structure of the proposed model

In the proposed model, we don't consider the sequence of objects, because the sequence is fairly influenced by saliencies of objects against the background. Furthermore, user could miss the object in his view and the proposed system has the purpose to help user's difficulty in finding a related object. In those cases, only combinations of objects of attention are considered importantly. If the intention consists of multiple actions, then the sequence of the objects would be considered to infer the composition of actions in the order. In this paper, we consider intentions of single actions only.

2.2 Deep Auto-encoder

Since I in Eq. (1) is the element of finite set of intentions, the problem, selection of intention which has the maximum affordance for given objects could be regarded as encoding from objects to the intention code.

The proposed model is constructed using deep auto-encoder based on the restricted Boltzmann machine (RBM) [15, 16]. Boltzmann machine is the stochastic machine which has a binary state with the probability. And restricted Boltzmann machine has the constraint that nodes are never connected with nodes in the same layer. By restricting the connections, more efficient learning algorithms can be used to train the network.

The neurons of Boltzmann machines has the stochastic behavior and the learning of the machine models the input patterns according to a Boltzmann distribution with assumption of network's thermal equilibrium.

Deep network structure can capture higher-order internal relationships or features which are unobservable directly. In the case of RBM, deep belief network stacked RBMs to construct deep hierarchical network. Although they are regarded as powerful high-level feature extracting machines, for a long time, deep network style learning machines are not practical because of its slowness on learning phase and absence of efficient learning algorithm. As G. Hinton et al. proposed the plausible pre-training and succeeding fine-tuning learning scheme, deep networks became popular structure for many machine learning fields.

Auto-encoder can be divided in to two parts, the encoder and the decoder. The encoder does transform data into relatively low-dimensional code while the decoder, the counterpart of the encoder, tries to recover the data with original dimension from the code. These are used to recognize intentions and reconstruct corresponding objects. As stated above, each node of RBM has a real-valued probability of the activation. At the code layer, these probabilities are evaluated and compared to decide which intention is most probable for the given input vector, as represented in Eq. (1). Subsequently, through the the decoder part, objects related to the intention are reconstructed and some objects are recommended as shown in Eq. (2). In later subsections, detailed explanations are presented for each step.

2.3 Intention Recognition

Since we are to recognize simple intentions without considering the sequence of objects of attention, we can build the visible (input) vector of RBM as the binary vector, positions of which represents whether the user saw those objects with attention. For instance, if 6 objects are considered, the dimensionality of input vector is 6. When the objects labeled 1 and 3 get user's attention, input vector becomes [1, 0, 1, 0, 0, 0]. We trained the encoder to learn intention codes corresponding to input patterns which represents a combination of objects of user's attention. And once the intention is predicted (i.e., intention code is produced in code layer), the related objects are decoded through the decoder part. Thus, intention to object and object to intention affordances can be modeled in encoder and decoder part of SDA, respectively.

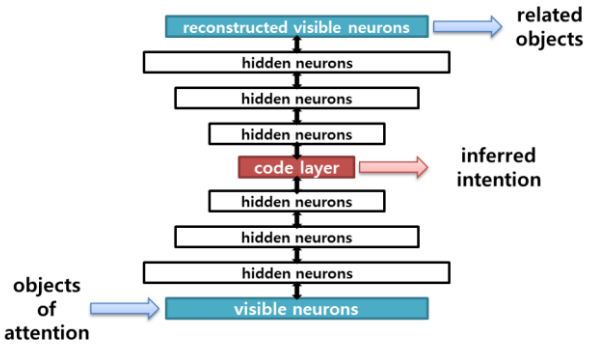

Fig. 2. The structure of deep auto-encoder

Fig. 2. shows the structure of deep auto-encoder used in the proposed model. At the bottom visible layer, objects of attention are represented by a binary vector and are given as input. And upper hidden layers extract statistical features hierarchically. Then at the code layer which is located in the middle, intention code is activated and these codes are used to classify intention and reconstruct the related objects at the uppermost layer. By adopting the deep auto-encoder, intention recognition and extraction of the object which has the highest affordance to predicted intention can be modeled in a natural way. And the SDA has the robustness against the noise also (for example there is a missing object) since it extracts higher-order features.

2.4 Recommendation of Object

The aim of recommendation is to suggest the unseen but important objects related to the specific intention. For

example, the user who wants to eat ramen, he might need to find ramen, a gas stove, and a pot. If his intention is predicted as the "eat ramen" and he seems to miss the a pot, the model recommends the pot based on intention-object affordance. Once the intention is predicted using the encoder, related objects are reconstructed through the decoder part of deep auto-encoder. By comparing the objects of attention with reconstructed objects, missing objects can be detected. Here we simply define $f(\cdot)$ of Eq. (2) as set operator \, where $f(X,Y)$ yields the relative complement of X in Y. The objects recommended are therfore calculated as:

$$\{obj_{recommended}\} = \{obj_{related}\} \backslash \{obj_{attention}\}, \quad (3)$$

where $\{obj_{attention}\}$ is reprenseted by the input and $\{obj_{related}\}$ is reconstructed output vector of the deep auto-encoder. So, using Eq. (3), objects which are much related with the intention but not be perceived enough by the user will be recommended.

3 Experiments

3.1 Experimental Setup

To show the performance of the proposed intention recognition and object recommendation, the experimental environment was set up. The scene of a kitchen containing 12 objects related to certain actions used for the experiments is shown in Fig. 3. The scene is the picture of a kitchen in which the objects are a ramen box, an instant coffee box, knives, an electric kettle, a pot, a gas stove, beer cans, mineral waters, glasses, breads, wines, and mugs. The user sees the scene displayed on the eye-tracker with his own intention.

Fig. 3. Exerimental scene and objects

In order to interpret the user's intention by understanding their attention for specific objects, we use Tobii 1750 eye-tracker. It has 17-inch TFT-LCD monitor and 1280x1024 resolution. The accuracy of the eye-tracker is 0.5 degrees, a sampling rate is 50 Hz and various gaze information could be obtained [17]. Fig. 4 is a picture of the Tobii 1750 eye-tracker.

Fig. 4. Tobii 1750 eye-tracker

Among six persons who participated in experiments, five person's data is used to train the model and remaining one person's data is used to test. Four different intentions such as 'eat noodle', 'drink beer', 'drink water' and 'eat bread'. are simulated using eye-tracker from the scene containing 12 objects.

Participants are instructed to look at the scene with specific intention for 10 seconds. During the experiments, the gaze information of users is gathered by eye-tracker system. This information consists of gaze path, gazing time on specific areas [9, 10]. If we assume that the gaze amount is proportional to user's attention level, we can binarize the user's attention on objects by some threshold. That is,

$$\boldsymbol{x} = [x_1, \ldots x_n],$$

where n is the number of objects in the scene $x_i = 1$ if gaze time on the *i*-th object exceeds the threshold θ and $x_i = 0$ otherwise. n is 12 in this experiment. This vector of objects attended is given as the input to the deep auto-encoder to predict the intention and recommend the objects that are unattended by the user.

The deep auto-encoder consists of 7 hidden layers which have dimensionalities 200, 100, 50, 15, 50, 100, 200, respectively. Lowermost 3 layers are used as the encoder, the middle 4th layer is used as representational layer of the intention inferred from the encoder, and highermost 3 layers are used as the decoder. The number of epochs for training is set to 100.

3.2 Experimental Results

Table I shows the accuracy of the intention recognition. As stated previously, five persons' data are used to train and one person's data are used to test. Task 1, 2, 3 and 4 in the table are 'eat noodle', 'drink coffee', 'drink beer' and 'eat bread' respectively.

Table I. Accuracy of intention recognition

	Task 1	Task 2	Task 3	Task 4
Train	100 %	86.5 %	100 %	100 %
Test	83.3 %	83.3 %	83.3 %	84.7 %

As shown in the table, test accuracies of recognition are over 80%, which means that the proposed intention recognition model is plausible to detect the user's intention based on eye gaze information. Since the purpose of the system is to recommend the object which is not perceived enough by the user but related his intention, the system should recognize the intention under that condition also. In other words, intention should be recognized even though there are some objects related to user's current intention but not marked as 1 in the input vector since the user couldn't find it. Table II shows the test performance of the case where there is a missing object. The performance shows the plausible recognition accuracy even though one object is missing.

Table II. Accuracy of intention recognition with a missing object

	Task 1	Task 2	Task 3	Task 4
Test	72.3 %	77 %	83.3 %	79.8 %

For recommendation of the object corresponding to a predicted intention, the root mean squared error (RMSE) of the test data is almost 0 (much lesser than 0.01) and 0.54 in the case that there is no missed object and one object is missing, respectively.

4 Conclusion

In this paper, we proposed the intention recognition as well as object recommendation system, and performed experiments on the environmental scenes of eye-tracker device. The model uses the deep auto-encoder as a main part for object affordance modeling. A missing object is successfully recommended by comparing the reconstructed results of the auto-encoder with user's attention.

Experimental section describes the generalization performance of the proposed intention recognition model and also shows meaningful accuracy of recommendation to provide interactive services.

In future works, we would implement the system on humanoid robot platform and obtain head-pose and intentional gestures to perform the joint attention. And incremental affordance learning would be considered to make the robot learn interactively. In this paper, the intention is composed of only one action, but for more general and natural HAI system, we will consider the intents consisting of several sequential actions.

Acknowledgement

This research was supported by the Industrial Strategic Technology Development Program (10044009) (50%) and the R&D program (10041826) (50%) funded by the Korea Ministry of Knowledge Economy (MKE) and the Korea Evaluation Institute of Industrial Technology (KEIT).

References

[1] Premack, D. and G. Woodruff.: Does the chimpanzee have a theory of mind?, Behavioral and Brain Sciences, Vol. 1, No. 4, pp. 515-526, (1978)

[2] Gibson, J. J.: The ecological approach to visual perception, Psychology Press, (1986)

[3] Kjellström, H., Romero, J., & Kragić, D.: Visual object-action recognition: Inferring object affordances from human demonstration, Computer Vision and Image Understanding, Vol. 115, No. 1, pp. 81-90, (2011)

[4] Montesano, Luis, et al.: Learning Object Affordances: From Sensory-Motor Coordination to Imitation, Robotics, IEEE Transactions on, Vol. 24. No. 1, pp. 15-26, (2008)

[5] Zhu, C., Cheng, Q., and Sheng, W.: Human intention recognition in smart assisted living systems using a hierarchical hidden markov model, IEEE International Conference on Automation Science and Engineering, pp. 253-258, (2008)

[6] Schrempf, O. C., and Hanebeck, U. D.: A generic model for estimating user-intentions in human-robot cooperation, In Proceedings of the 2nd International Conference on Informatics in Control, Automation and Robotics, Vol. 5, (2005)

[7] Kelley, R., Wigand, L., Hamilton, B., Browne, K., Nicolescu, M., and Nicolescu, M.: Deep networks for predicting human intent with respect to objects,

Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction, pp. 171-172, (2012)

[8] Hwang, B., Jang, Y. M., Mallipeddi, R., and Lee, M.: Probabilistic human intention modeling for cognitive augmentation, IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp. 2580-2584, (2012)

[9] Rish, I.: An empirical study of the naive Bayes classifier, IJCAI 2001 workshop on empirical methods in artificial intelligence, Vol. 3, No. 22, pp. 41-46, (2001)

[10] Vertegaal, R., Shell, J. S., Chen, D., and Mamuji, A.: Designing for augmented attention: Towards a framework for attentive user interfaces, Computers in Human Behavior, Vol. 22, No. 4, pp. 771-789, (2006)

[11] Jacob, R. J.: What you look at is what you get: eye movement-based interaction techniques, Proceedings of the SIGCHI conference on Human factors in computing systems, pp. 11-18, (1990)

[12] Vertegaal, R., Slagter, R., van der Veer, G., and Nijholt, A.: Eye gaze patterns in conversations: there is more to conversational agents than meets the eyes, In Proceedings of the SIGCHI conference on Human factors in computing systems, pp. 301-308, (2001)

[13] Jang, Y. M., Mallipeddi, R., Lee, M., Lee, S., and Kwak, H. W.: Human implicit intent transition detection based on pupillary analysis, International Joint Conference on Neural Networks (IJCNN), pp. 1-7, (2012)

[14] Jang, Y. M., Lee, S., Mallipeddi, R., Kwak, H. W., and Lee, M.: Recognition of human's implicit intention based on an eyeball movement pattern analysis, In Neural Information Processing, pp. 138-145, (2011)

[15] Hinton, G. E., and Ruslan R. Salakhutdinov.: Reducing the dimensionality of data with neural networks, Science, Vol. 313, 5786, pp. 504-507, (2006)

[16] Hinton, G. E., Osindero, S., and Teh, Y. W.: A fast learning algorithm for deep belief nets, Neural computation, Vol. 18, No. 7, pp. 1527-1554, (2006)

[17] Eye tracking system of Tobii technology, http://www.tobii.com/