Description and recall of the object using processes with scene change using LSTM

Masaki Yano[1,a], Tadashi Matsuo[1] , Nobutaka Shimada[1]

***Abstract*— When people manipulate an object, they have some intention to change the *state* of the object itself or scene environment. In many cases object operations consist of some steps with state changes and are described as a series of such state transitions of both the object and the scene.**

In this paper, we propose a method that describes and recalls co-occurrence of a human action and a scene change when a person uses an object. The person's skeleton is used as a human posture feature and a scene depth image is used as scene feature. We train a Long Short-Term Memory(LSTM)[1] model from the human posture and scene features. The model recalls the person's action and scene change from a current and a goal state of the person and the scene as a series of the co-occurring human action and scene transition. We evaluate the availability of the proposed method by an experiment in which a person uses a chair.

I. INTRODUCTION

Recently, autonomous robots work in the various fields and they will support us more at living space in the future. These robots need to manipulate objects which we usually use. However, there are many kinds of objects in daily life, so it is very hard to develop softwares which make the robots manipulate any objects.

One of the solutions to this problem, there is a method called "Learning from Demonstration(LfD)[2]". LfD is the research that a robot watches human's action and learns the action. There are a lot of research regarding LfD[3],[4], however, few researches consider that the robot's action when the scene state doesn't change in an expected way.

Our research tries to considers both the robot's action and the scene or object's state change caused by the action. Specifically, the robot confirms that the scene state changes into the expected or not. The robot goes to the next process when the scene state expectedly has changed. If not, the robot generates actions that can change the scene state into the expected one.

In this paper, we propose a method that describes and recalls co-occurrence of a human action and a scene change when a person uses an object. We train a LSTM model from the human posture features and the scene features. The model can recall the human action and scene change which corresponds to the current state. We aim that the robot learns the human's action by using this framework.

II. DESCRIBING AND RECALLING CO-OCCURRENCE OF THE HUMAN ACTION AND THE SCENE CHANGE

We train a LSTM model from the human action and the scene change. Because the model learns co-occurrence of a human posture feature and a scene feature, the model can recall the expected, namely trained, next posture and scene features based on the current input features, which is the next step forward to a specified goal scene of object use. The person's skeleton is used as a human posture feature and a scene depth image is used as scene feature. At this time, we don't consider the model inputs complementarity.

A. Describing the Scene Feature by Using Sparse Auto-Encoder

We use Sparse Auto-Encoder[5] for describing the scene depth image as the low dimensional vector.

Fig. 1 shows the examples of training data for learning the Auto-Encoder. The training data includes the seven kinds of scene state which are shown Fig. 1.

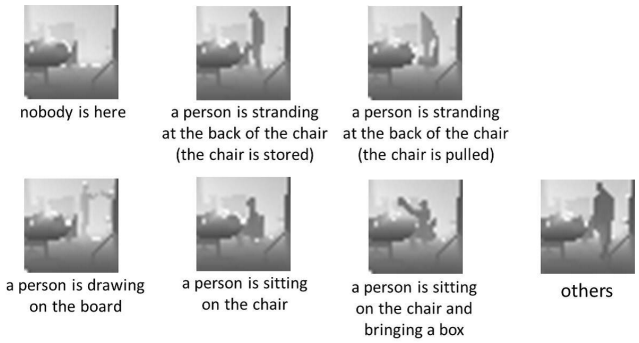

Fig. 1. The examples of the training data for learning the Auto-Encoder model

Fig. 2 shows the Auto-Encoder which we build. The input of the Auto-Encoder is a scene depth image(32x32) and we use the output of middle layer (the result of encoding) as the scene feature.

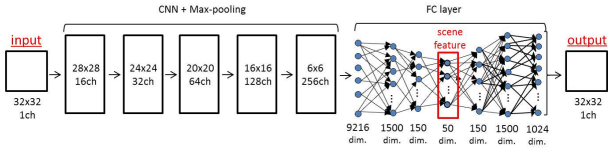

Fig. 2. Network structure of the Auto-Encoder

B. Describing the Human Action and the Scene Change by Using LSTM

Fig. 3 shows the examples of training data for learning the LSTM model. The training data includes the six kinds of human action which are shown Fig. 3.

[1]Ritsumeikan University
[a]yano@i.ci.ritsumei.ac.jp

Fig. 3. The examples of the training data for learning the LSTM model

Fig. 4 shows the LSTM model which we build. The model's inputs are the current skeleton, the current scene, the goal skeleton, the goal scene and the time until goal. The model's outputs are the next frame skeleton and scene. The model's hidden layer is built with 2 layers and each layer has 600 LSTM nodes. We train this LSTM model with the training data shown Fig. 3. The loss function to minimize is shown equation (1).

$$
\begin{aligned}
E(t) &= w_1\left(s_{t+1} - y_{t+1}^{(0,1,\ldots,t)}\right)^2 + \frac{w_2}{N-2}\sum_{t=2}^{N-1}\left(s_{t+i} - y_{t+i}^{(0,1,\ldots,t)}\right)^2 \\
&+ w_3\left(s_{t+N} - y_{t+N}^{(0,1,\ldots,t)}\right)^2 + w_4\left(b_{t+1} - z_{t+1}^{(0,1,\ldots,t)}\right)^2 \\
&+ \frac{w_5}{N-2}\sum_{t=2}^{N-1}\left(b_{t+i} - z_{t+i}^{(0,1,\ldots,t)}\right)^2 + w_6\left(b_{t+N} - z_{t+N}^{(0,1,\ldots,t)}\right)^2 \\
&+ \frac{w_7}{N-1}\sum_{t=0}^{N-1}\left(z_{t+i}^{(0,1,\ldots,t)} - z_{t+i+1}^{(0,1,\ldots,t)}\right)^2 \qquad (1)
\end{aligned}
$$

where w_1 to w_7 is the normalization coefficients, $y_{t+i}^{(0,1,\ldots,t)}$ is the human skeleton of recall at time is $t+i$, $s_{t+i}^{(0,1,\ldots,t)}$ is the actual (namely observed) human skeleton at the time. $z_{t+i}^{(0,1,\ldots,t)}$ is the recalled scene when the time is $t+i$, $b_{t+i}^{(0,1,\ldots,t)}$ is the actual scene at the time.

C. Recalling the Human Action and the Scene Change

Fig. 5 shows the result of recalling by the model. The upper row shows the result of recalling and the lower row shows the actual observation. The initial state is the state that a person is standing near a chair, and the goal state is the state that the person is sitting on the chair. The model successfully recalled a sequence of the person's actions in which the person sits down on the chair after he pulled the chair.

Compared with the actual observation, the start timeing of each action is different, however, the order in the realized actions is the same.

III. CONCLUSION

In this paper, we propose a method that describes and recalls co-occurrence of a human action and a scene change when a person uses an object. We built the LSTM model and trained it with the human skeletons and the scene features.

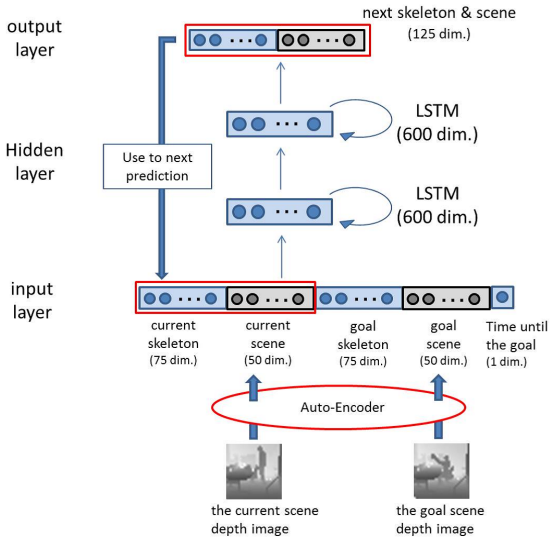

Fig. 4. Network structure of the LSTM

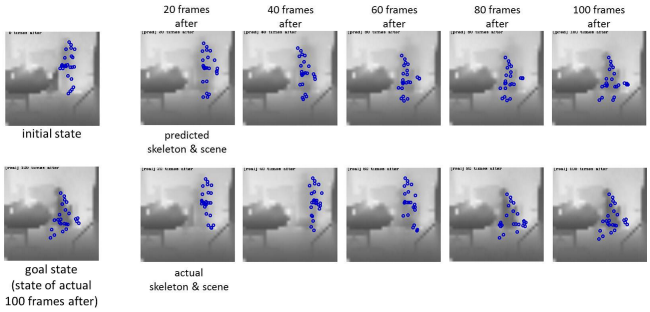

Fig. 5. The result of the human actions and the scene change recalling; Blue marks are the person's skeleton

In the prediction experiment, the model recalled them the same order as actual.

We are extending to allow our proposed method to work even if the human fails to change the scene state. For example, when the person failed to pull the chair, the model recalls the action that the person pulls the chair again.

ACKNOWLEDGMENT

This work was supported by JSPS KAKENHI Grant Numbers JP24500224, JP15H02764.

REFERENCES

[1] Hochreiter, Sepp, Jürgen Schmidhuber. "Long short-term memory", Neural computation 9.8 (1997): pp. 1735-1780.
[2] Atkeson, Christopher G., and Stefan Schaal. "Robot learning from demonstration.", ICML. Vol. 97. pp. 12-20.
[3] Lee, Kyuhwa, et al. "A syntactic approach to robot imitation learning using probabilistic activity grammars.", Robotics and Autonomous Systems 61.12 (2013): pp. 1323-1334.
[4] Sermanet, Pierre, et al. "Time-Contrastive Networks: Self-Supervised Learning from Multi-View Observation.", arXiv preprint arXiv:1704.06888 (2017).
[5] Tadashi Matsuo, Nobutaka Shimada. "Construction of Latent Descriptor Space of Hand-Object Interaction",The 22nd Joint Workshop on Frontiers of Computer Vision (FCV2016): pp. 117-122.