AFRL-AFOSR-JP-TR-2016-0013

Predictive Coding Strategies for invariant object recognition and volitional motion control in neuromorphic agents.

Dae-Shik Kim
Korea Advanced Institute of Science and Technology

09/02/2015
Final Report

Air Force Research Laboratory
AF Office Of Scientific Research (AFOSR)/ IOA
Arlington, Virginia 22203
Air Force Materiel Command

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
21-12-2015	Final	24-09-2012 to 23-09-2015

4. TITLE AND SUBTITLE
Predictive Coding Strategies for invariant object recognition and volitional motion control in neuromorphic agents.

5a. CONTRACT NUMBER
FA2386-12-1-4089

5b. GRANT NUMBER
Grant 12RSZ109_124089

5c. PROGRAM ELEMENT NUMBER
61102F

6. AUTHOR(S)
Dae-Shik Kim

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Korea Advanced Institute of Science and Technology
291 Daehak-ro, Yuseong-gu
Taejon, 305701 KR

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
AOARD
UNIT 45002
APO AP 96338-5002

10. SPONSOR/MONITOR'S ACRONYM(S)
AFRL/AFOSR/IOA(AOARD)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
12RSZ109_124089

12. DISTRIBUTION/AVAILABILITY STATEMENT
Distribution Code A: Approved for public release, distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT
Aim #1: Learning invariant representations of environments through experience has been important area of research both in the field of machine learning as well as in computational neuroscience. In this study, we employed a novel method for the discovery of invariants from a single video input based on the learning of the predictability of spatio-temporal relationships between inputs. To this end, videos containing spatio-temporal movements of unlabeled natural objects were used.

Progress:
1) Conducted real-time invariant perception and tracking of natural images
2) Conducted real-time invariant perception and tracking of video objects

Aim #2: Volitional movements are a hallmark for human behavior. In this project, we hypothesized that visual memory of past motion trajectories may be used for selecting future behavior. In other words: following free energy principle, apparent volitional movements can be generated by minimizing the difference between what the agent expected to see and what it effectively sees.

Progress:
1) Tested of robotic systems prediction-based pseud-volitional movements in a complex environment requiring adaptive modifications. Phase I: physics modeled in Gazebo-type of simulated environment
2) Tested of robotic systems prediction-based pseud-volitional movements in a complex environment requiring adaptive modifications. Phase II: tested in a real physical environment

15. SUBJECT TERMS
Artificial Intelligence, Cognitive Science

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Brian Lutz
Unclassified	Unclassified	Unclassified	SAR	17	19b. TELEPHONE NUMBER *(Include area code)* 315-227-7006

Final Report for AOARD Grant AF2386-12-1-4087

"**Predictive Coding strategies for invariant object recognition and volitional motion control in neuromorphic agents**"

09/01/2015

Name of Principal Investigators (PI and Co-PIs): KIM DAE SHIK
- e-mail address : daeshik@kaist.ac.kr
- Institution : Korea Advanced Institute of Science and Technology
- Mailing Address : 34141
- Phone : +82-42-350-3490
- Fax : +82-42-350-8170

Period of Performance: 09/24/2012 ~ 09/23/2015

Abstract: In this project, we developed and proposed brain-like neuronal models. Deep learning approaches such as deep convolutional neural networks (CNNs) and recurrent neural networks (RNNs) have recently outperformed in various research areas. Moreover, they inspired by computational mechanisms of human brains as a neural network model. Hence, we conducted experiments with different tasks to solve real-world problems, and proposed models in this work were based on the deep learning approach. As a result, we reached state-of-art performances in some of the visual tasks, and we found a meaningful relationship between a computational models and human developmental process.

Introduction

The main objective of this research project is to propose brain-like computational models to be able to recognize visual object robustly (invariant with rotation and transition) and generate goal-directed motor actions. Deep learning is a family of algorithms inspired by the human brain, and it recently shows the state-of-the-art performance in various fields of studies related to vision, speech, and language processing. We have studied and developed deep-learning-based models for extracting feasible feature for such tasks, which share the core concepts with predictive coding strategies. Unlike classical computer vision approaches, deep learning approach learns meaningful feature hierarchies in an autonomous manner just like the brain does. By utilizing these invariant features learned by deep architectures, we successfully developed neuronal models and obtained results with following tasks.

Video scene understanding: A computational model for scene understanding was proposed based on deep convolutional neural networks to improve recognition accuracy.

Facial expression recognition: A deep-learning-based model for facial expression recognition was formulated. It could recognize emotional status of people regardless of background conditions and classify seven different classes of expression.

Robust real-time object tracking: A framework for robust real-time object tracking was proposed. In this framework, a deep CNN based object recognition algorithm was combined with a conventional visual tracker and detector.

Human action recognition: A deep temporal CNN based model was proposed to recognize human actions based on time-series of pose features.

Learning for goal-directed actions using RNNPB: A robotic experiment in the virtual workspace was conducted with a robotic agent equipped with RNNPB model which is a kind of recurrent neural network model to be able to recognize and generate multiple temporal dynamics. The robotic agent showed human-like developmental dynamics during training goal-directed motor actions with an imitation learning task.

Scene Understanding

Introduction

The goal of this research is to improve the video scene understanding rate. This could be used to help to improve performances of other research areas such as human action recognition by making use of the relationship between specific scenes and actions. For this purpose, we formulate our model for scene understanding based on deep convolutional neural networks.

Experiments

Datasets

We reported results on the popular Place205[1] database and MIT indoor 67 dataset(2) which has 2.5 million training images and 5,360 training images respectively. Place205 database has 205 scene categories, and MIT indoor 67 dataset has 67 indoor scene categories.

Training Procedure

We used Deep CNN model, especially Alexnet which consists of 5 convolution layers and three fully connected layers. Our experiment used training/fine-tuning splits. In training split, Deep CNN model except the classification layer was trained by using Place205 database and in fine-tuning split, only the last classification layer was trained by using MIT indoor 67 dataset.

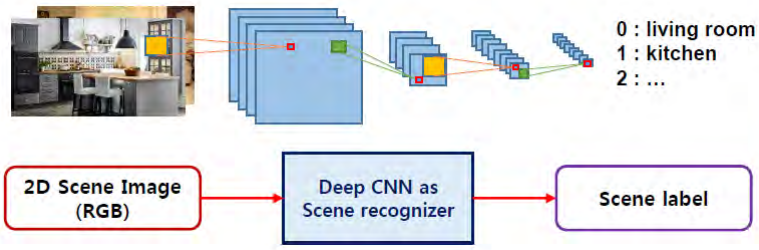


Figure 1. Our Alexnet model for scene understanding.

Results and Discussion

We report results on the Place205 database and MIT indoor dataset in Table1 and evaluate our experiment with the frequently used Top-1 and Top-5 results. We obtain best results compared to other methods with the model which is trained by Place205 database and is fine-tuned by MIT indoor dataset. From this research we demonstrate that our approach exceeds the current state-of-the-art methods.

Table 1. Top-1 and Top-5 compared to other methods. This table shows our method outperforms other research results.[3,4]

	Model	Top-1	Top-5	Data size	# Class	Model Size
Place205 database	Place-CNN (Zhou et al., 2014)	50.5756%	**80.9414%**	20,500	205	221M
	(our)Place-CNN	**50.5854%**	80.9122%			
	(our)Place-NiN	50.1219%	80.039%			26M
MIT indoor dataset	CNN-SVM (Razavian et al., 2014)	69.0%	X	1,340	67	X
	Place-CNN hybrid (Zhou et al., 2014)	70.80%	X			221M
	Place-NiN + MIT-indoor	**72.3134%**	93.6567%			26M

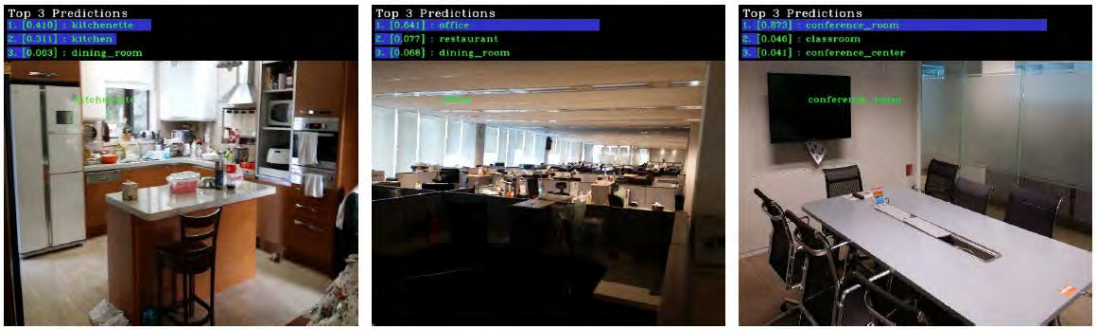


Figure 2. Examples of scene prediction with Place205 database

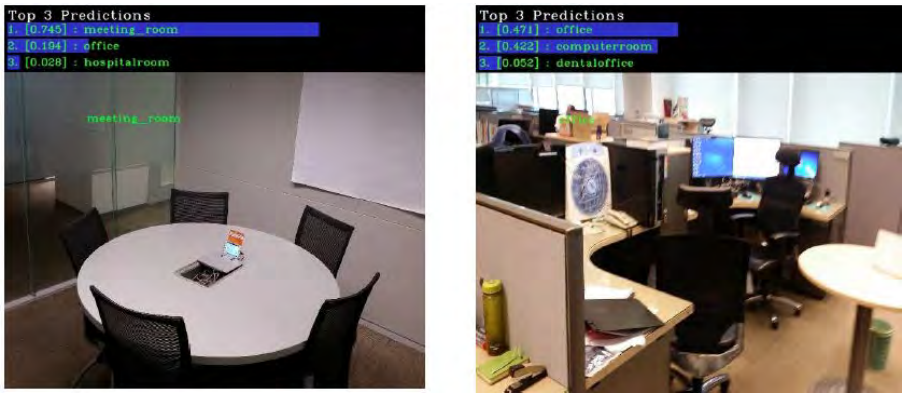


Figure 3. Examples of scene prediction with MIT indoor dataset

References

1. B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. “Learning Deep Features for Scene Recognition using Places Database.” Advances in Neural Information Processing Systems 27 (NIPS), 2014
2. A.Quattoni, and A.Torralba. “Recognizing Indoor Scenes.” IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2009.
3. Zhou, Bolei, et al. "Learning deep features for scene recognition using places database."Advances in Neural Information Processing Systems. 2014.
4. Razavian, Ali S., et al. "CNN features off-the-shelf: an astounding baseline for recognition." Computer Vision and Pattern Recognition Workshops (CVPRW), 2014 IEEE Conference on. IEEE, 2014.

Facial Expression Recognition

Introduction

Humans emotional status mainly appears in their facial expression. If we know that information, we can react differently according to the target's status. The specific aim of facial expression recognition is searching human face and classifying target's emotional status by deep convolutional neural network model. Classification is performed with seven different classes of expression: Angry, disgust, fear, happy, sad, surprise and neutral.

Experiments

Facial expression recognition task consists of three-steps: a) face detection, b) face tracking, c) expression recognition. Face detection and tracking is performed with hand-crafted model, and expression recognition is performed with deep learning based classification technology.
Experiment was performed with Kaggle facial expression recognition challenge dataset with following specification:

- 48x48 grayscale images of faces
- 7 classes
- Training set: 28,709
- Validation set: 3,589
- Test set: 3,589

Figure 4. Kaggle facial expression recognition challenge dataset

We used Caffe CNN model for realizing our classification model. Specific model is described on following figure.

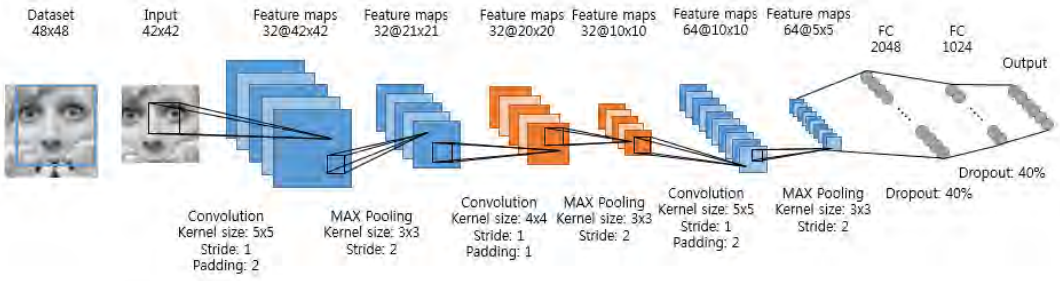


Figure 5. CNN architecture used for classification

We trained the model with training set and enhance the performance with several parameter tuning.

Results and Discussion

The highest accuracy we get from parameter tuning was average 70.74%. Individual accuracy due to each class is shown on following table. This result is slightly lower than Kaggle challenge winner's record (71%).

Table 2. Confusion matrix of trained CNN model

	angry	disgust	fear	happy	sad	surprise	neutral
angry	61.1	1.0	6.3	3.5	15.3	1.6	11.2
disgust	23.6	67.3	1.8	0.0	1.8	1.8	3.6
fear	10.6	0.2	49.5	3.0	19.7	8.1	8.9
happy	1.4	0.0	0.8	88.7	3.9	1.6	3.6
sad	6.1	0.0	8.1	5.1	65.2	0.8	14.8
surprise	2.2	0.0	5.8	3.6	2.9	83.2	2.4
neutral	3.6	0.3	3.7	7.3	14.4	1.5	69.2

References

1. Tang, Yichuan. "Deep learning using support vector machines." CoRR, abs/1306.0239 (2013).
2. Bergstra, James, and David D. Cox. "Hyperparameter Optimization and Boosting for Classifying Facial Expressions: How good can a" Null" Model be?." arXiv preprint arXiv:1306.3476 (2013).
3. Kaggle Facial Expression Recognition challenge Database, https://www.kaggle.com/c/challenges-in-representation-learning-facial-expression-recognition-challenge
4. Jia, Y. (2013). Caffe: An open source convolutional architecture for fast feature embedding.

Robust real-time object tracking

Introduction

This research showed the framework of robust long-term and real-time tracking of an object recognizing what it is. Here recognizing means classification for the object based on image database using the convolutional neural network (CNN). Tracked object a classifier corresponding to the object is trained on-line using positive and negative constraints inside a local region. A detector module is integrated into the framework to overcome the disappearing problem. Proposed framework selects the best result from several independent components and estimates the error at the same time. Kalman Filter and Particle Filter are used inside filtering component to predict possible positions of the object in the next frame. We also use the CNN trained using ImageNet dataset (1000-classes). After that CNN classify the object region at every frame moment.

Experiments

In case that traditional tracking object on the long-term video stream, it using just feature vector from object region and doesn't concern what it is. It only focuses on the best feature for frame-to-frame bounding-box. In this study, we propose an advanced system that combines object tracking and recognition. It can autonomously track the target object without any supervision.

Using optical flow tracker and multi-scale detector on the Tracking-Learning-Detection platform, it can tracking the initialized object frame-to-frame even if the object moves out of camera view and come back inside. It was using positive/negative precision and positive/negative recall value for making error cancelation matrix. So this is the bootstrapping of the semi-supervised learning and then can update detector.

In this research also use CNN for classify the tracked object. We training the CNN based on ImageNet database which structure is Network in Network.

Finally, this system can tracking the initialized object at every frame and pass it through CNN for classification. It can run dual system simultaneously in real-time.

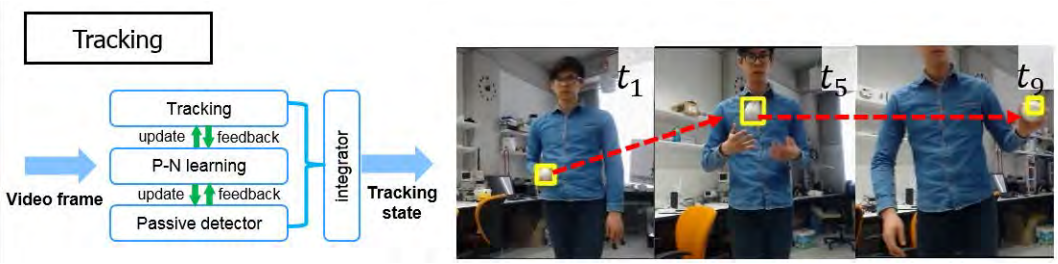


Figure 6. Tracking process and snapshots of result on ball-moving video

Convolutional neural network structure

- 4 convolution layers
- 4 pooling layers
- Using ReLu function
- Affect NIN structure
  ● 1x1 convolution kernel between each convolution layer
  ● Average max classifier

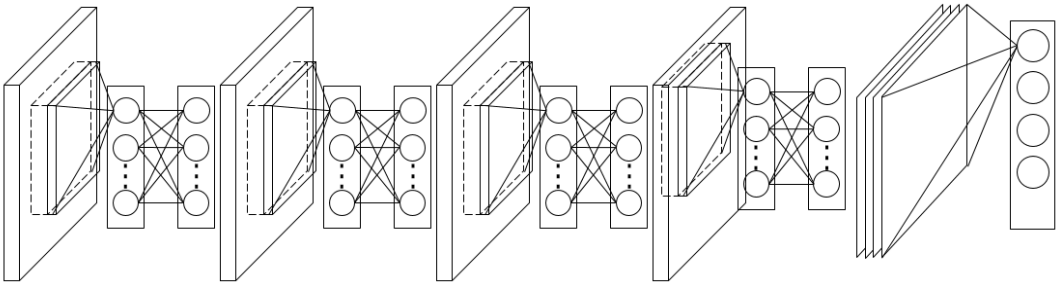


Figure 7. Network in Network structure for ImageNet classification

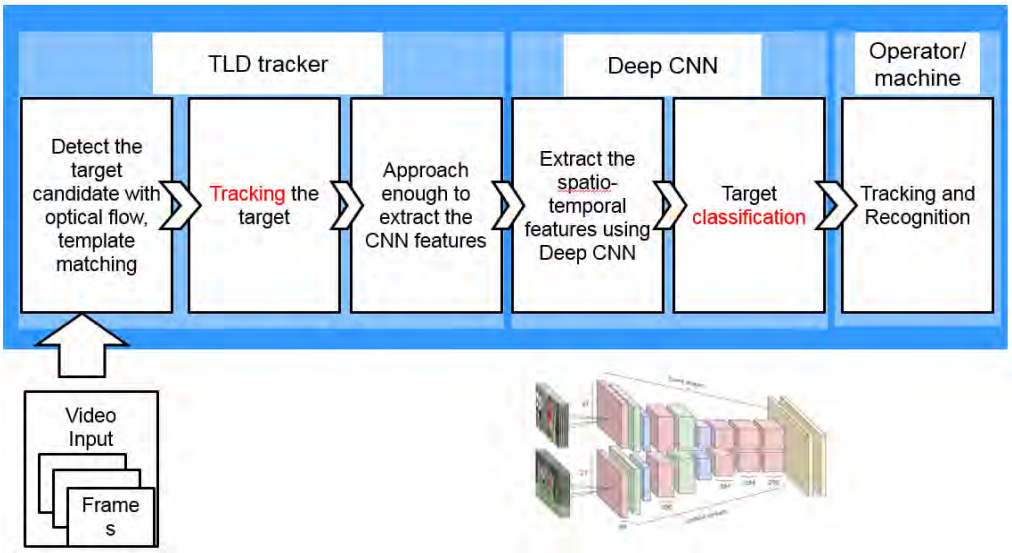


Figure 8. Integrated system for object tracking and recognition in real-time

Results and Discussion

As above pictures about mouse moving, this system can tracking not only the scale-variable object and also re-tracking object which moved out of camera and reappear inside of. In addition, it can recognize the object by classifier through CNN.

Recognition accuracy is 83% for Top-5 classes, and tracking performance is composed of precision and recall (precision – 68~80%, recall – 69~85%). Tracking performance suggest average value of each parameter from TLD video dataset.

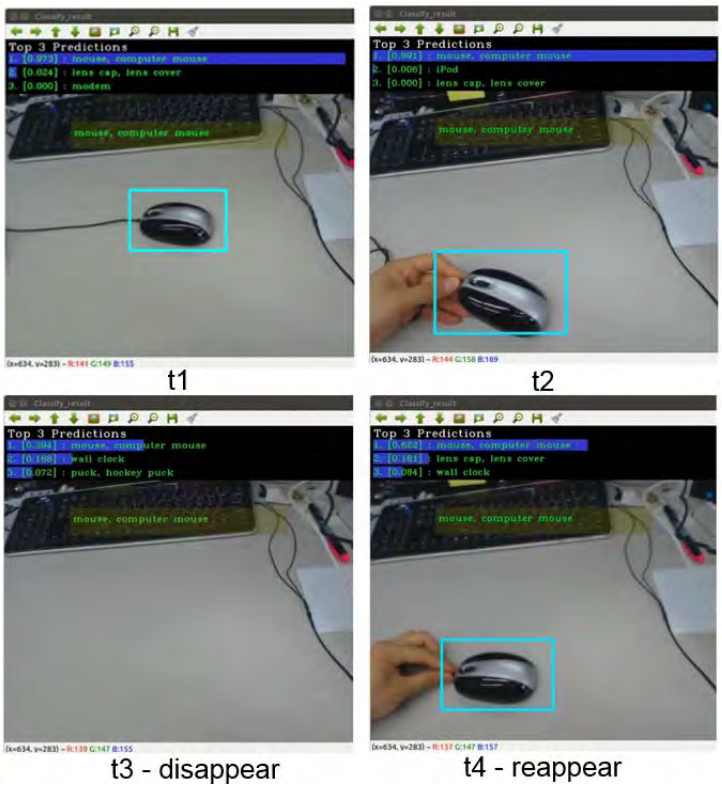


Figure 9. Result test about mouse moving video.

❖ Recognition performance

Num of classes	1000
accuracy	Top-5 83%

❖ Tracking performance

FPS	10~20
Precision	68~80%
Recall	69~85%

❖ Recall : precision object frames/true object frames
❖ Precision : true positive frames/recall frames

Reference.

1. Min Lin, Qiang Chen, Shuicheng Yan. "Network in Network"arXiv. *Mar 2014 v3.*
2. Z. Kalal, K. Mikolajczyk, and J. Matas. "Tracking-learning-detection." Pattern Analysis and Machine Intelligence, IEEE Transactions on, 34(7):1409 –1422, july 2012.

Human Action Recognition with Deep Neural Network

Introduction

Recognizing human actions from video data is one of the main characteristics toward to develop brain-like algorithms. In this study, a deep neural network model is proposed to classify human actions when time-series of 3D human poses are given. As extracting spatiotemporal motion features of human action, a hierarchal network architecture shows brain-like information process.

Experiments

Dataset

Recognizing human action is a wide and ambiguous definition of the problem because there are many possible applications (e.g. video surveillance, selecting sports highlights, hand gesture detection, and so on). MSR Kinect datasets are one of these variations. We choose Kinect datasets because they contain three-dimensional human pose data which could be easily obtained by using Kinect to test our model in the real world environment (see Figure 10).

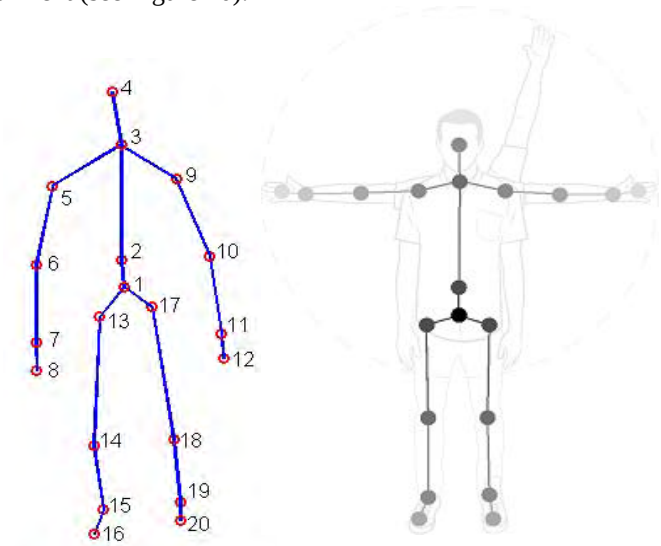


Figure 10. Skeleton and corresponding joint number of Kinect

MSR Online Action Dataset [(1)] is one of MSR Kinect datasets. It contains seven indoor action classes (Drinking, Eating, Using Laptop, Reading cellphone, Making phone call, Reading book, and Using remote). As it contains several subsets with different subjects and environments, it could be used as two-fold validation and tested as cross-environment action recognition.

Table 3 Subsets of the MSR Online Action Dataset

	S1	**S2**	**S3 DIFFERENT ENVIRONMENT**	**S4 LONG VIDEO**	**S0 NEGATIVE ACTION**
ACTIONS	7 (0~6)	7 (0~6)	7 (0~6)	1 (8)	1 (10)
SUBJECTS	8 (1~8)	8 (9~16)	8 (17~24)	12 (25~36)	3 (1, 5, 37)
TRIAL	2 (0~1)	2 (0~1)	2 (0~1)	3 (11~14)	5 (20~24)
TOTAL	**112**	**112**	**112**	**36**	**14**

Preprocessing

In this dataset, pose data is a time-series of multi-dimensional vectors which consist of three-dimensional position vectors of each joint as a real-world coordination. However, these are not invariant toward translation and rotation of human actors. Hence, we converted joint position vectors into the quaternion coordination. Time series of quaternion vectors were split with a sliding time window of which width was fixed.

Network architecture

A deep convolutional neural network (deep CNN) model was used as an elementary building block because it has showed an outstanding performance in the object classification task of the computer vision research area. In this works, temporal convolution units which extract temporal features were organized as a three layers accompanied with subsampling layers. Fully-connected layers with a softmax layer which are same with conventional neural network models finally classify actions based on extracted features from the temporal convolution units. As the inputs of the networks were time-windowed, classification was conducted for each time-windows.

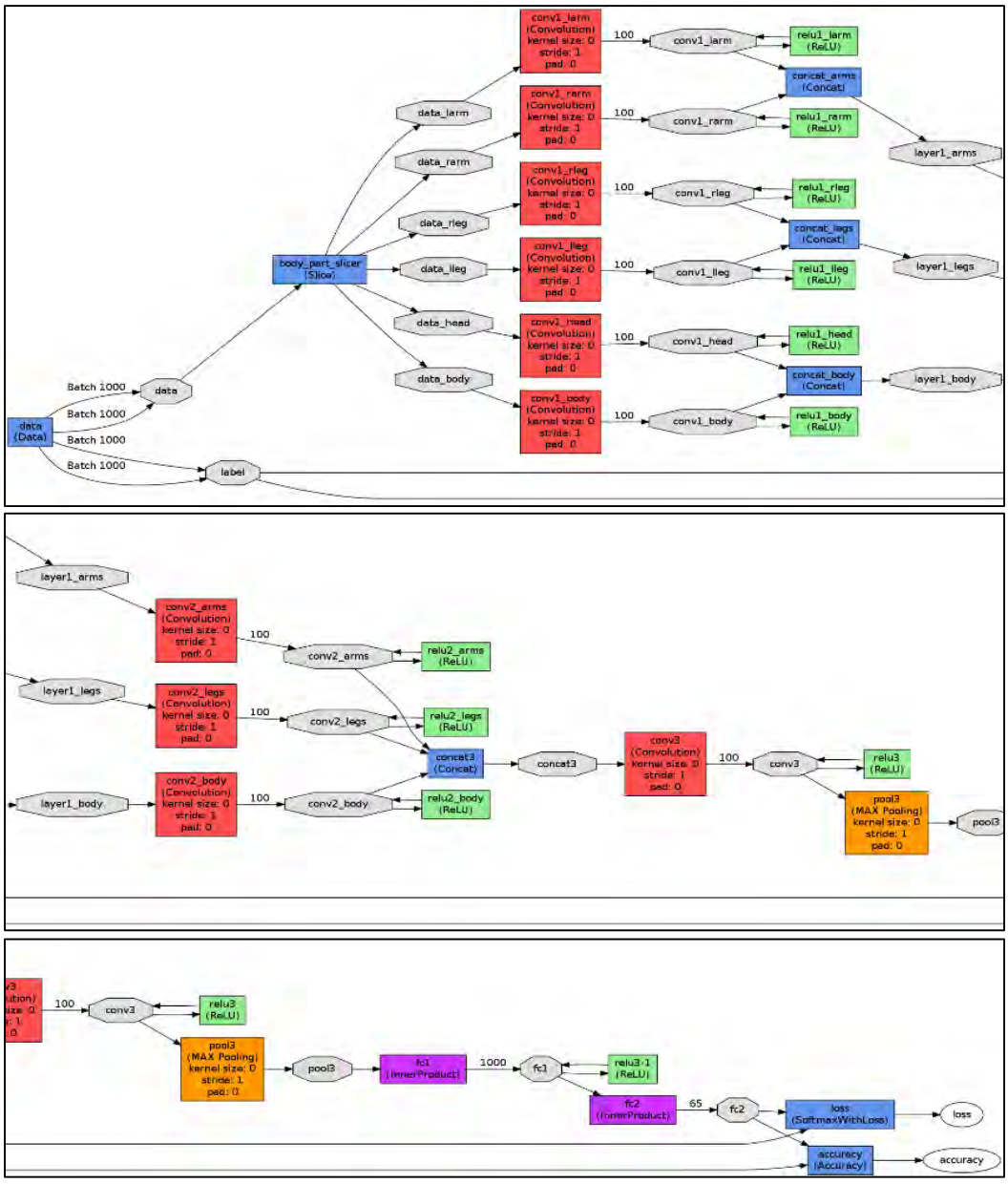


Figure 11. Network architecture of proposed model.

Result and Discussion

The proposed model was trained for 9,000 iterations with the subset S1 (see Table2). Recognition accuracies were measured by comparing with ground through in the subset S2. As a result, our model showed 50.89% of average recognition accuracy. According to a confusion matrix from the results (See Figure X), the class 'eating' was well whereas some classes such as 'using laptop' were not.

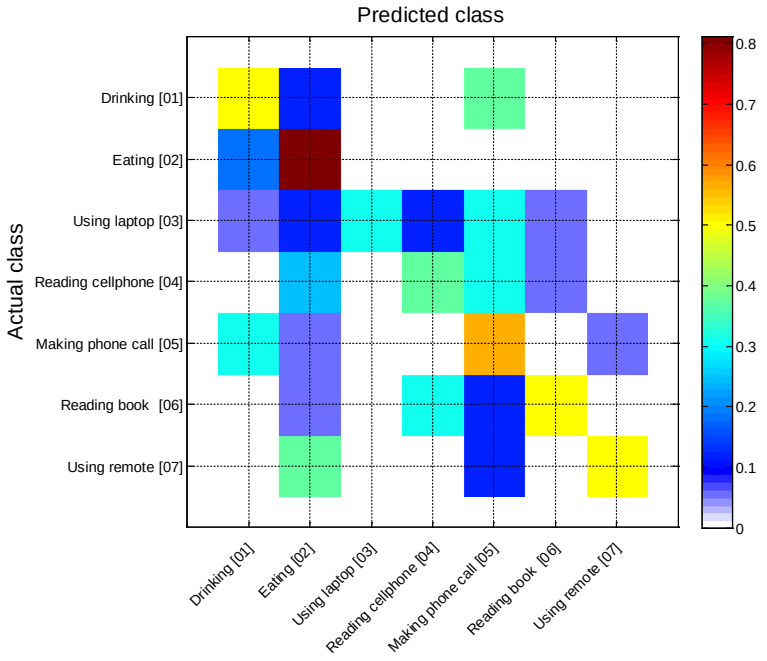


Figure 12. Confusion matrix.

The main reason of this variation between classes is that the proposed model used only short periods of time windows to determine actions. Hence, for next steps, we will adopt recurrent neural networks (RNNs) as combining with the deep CNN as a feature extractor. RNNs have been successfully implemented to model long-term time series. Moreover, a recent study about a language modeling with the RNN[(2)] showed that they could detect hierarchical temporal dynamic.

References

1. Gang Yu, Zicheng Liu, Junsong Yuan, “Discriminative Orderlet Mining For Real-time Recognition of Human-Object Interaction”, ACCV 2014.
2. Hermans, M., & Schrauwen, B. (2013). Training and analysing deep recurrent neural networks. In Advances in Neural Information Processing Systems (pp. 190-198).

Learning for goal-directed actions using RNNPB

Introduction

Imitation learning is also an important feature to develop brain-like algorithms as humans learn new actions by mimicking others. Developmental studies have showed that infants develop goal-directedness of sensory-motor actions. In the study of Carpenter et al.[1] young infants tend to ignore less salient properties as imitating only salient property of actions whereas older infants imitated both. In this study, we proposed a computational model to be able to learn multiple goal-directed motor behaviors in a robotic environment and investigated developmental dynamics of the proposed model while learning.

Experiments

Recurrent neural network with parametric bias (RNNPB) model[2] has an ability to memorize and regenerate multiple time-series. During error optimization process such as back-propagation through time (BPTT)[3], these multiple time-series are parameterized by a specialized neural units which name is parametric bias (PB) unit. Similar to the developmental study[1], we designed a robotic experiment in a simulation environment. In this experiment, a virtual robotic arm which consists of two joints moves in the two-dimensional workspace (See right parts of Figure 13).

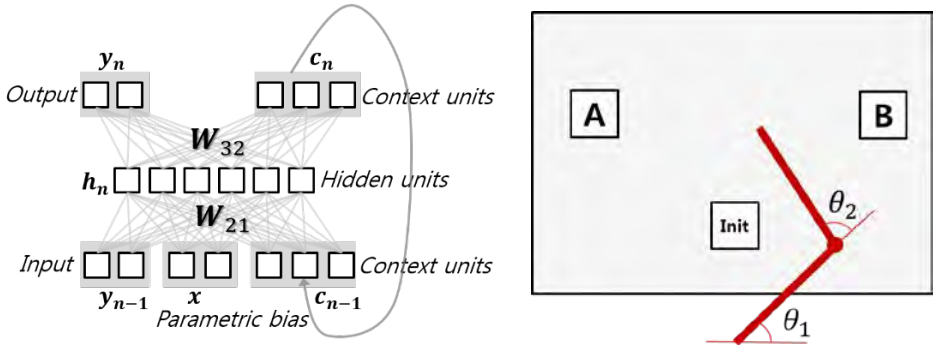


Figure 13. ***(left)*** **Architecture of RNNPB model consisting of three layers: input layer, hidden layer, and output layer.** ***(right)*** **Robotic arm moving from initial position to two different goal positions.**

As described below, six different motor behaviors (two separated goal positions with three different styles of movements) were defined to be trained. Proposed actions have two goal properties. The first one is related to the goal position, and it is called as 'the goal.' The second one is related to ways to reach to the goal positons, and it is called as 'the means.'

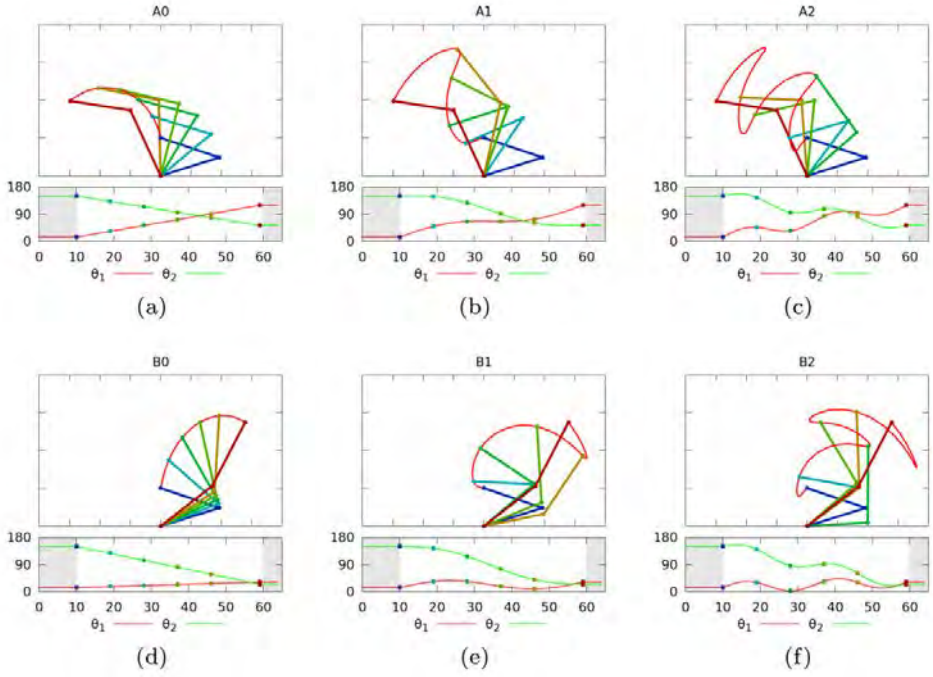


Figure 14. Three different movements for two goal positions. (a)-(c) reaches goal A, and (d)-(f) reaches goal B.

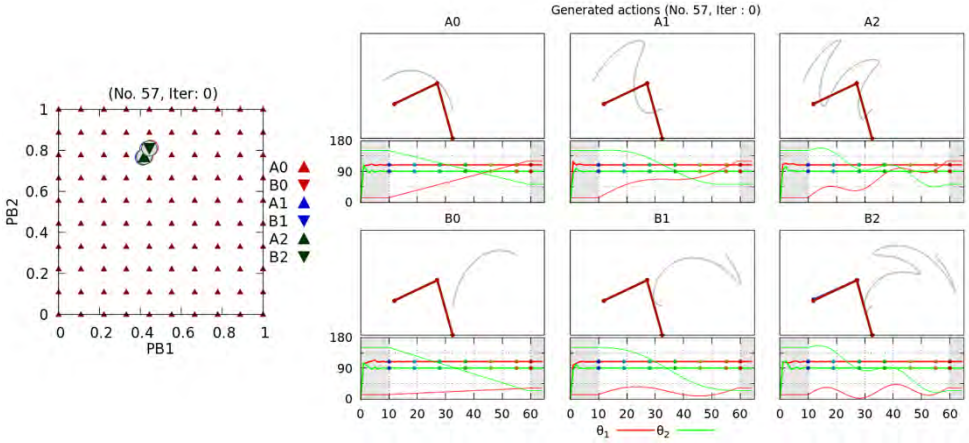
Results and Discussion

To examine how an ability of a robotic agent about imitating goal-directed actions is developed, the robotic agent equipped with the RNNPB model had been analyzed for every snapshot during training. Three snapshots were choose based on separation of the PB units, and analysis were conducted. For each snapshot, the robotic agent experienced the desired action as changing its internal status with corresponding PB values, and then it generated its own actions.

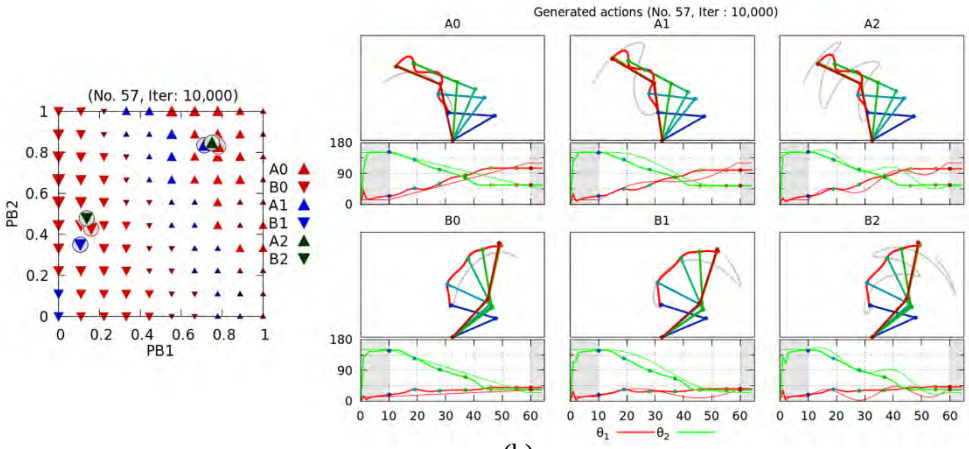
As a result, the agent showed staged development of its ability. When it is not trained yet, it did not produce the desired actions. It started to recognize and generated desired actions when the agented trained. PB unit values were also separated into the two parts. However, it was not followed styles of movements yet. When the agent was trained enough, it successfully recognized and generated for all of six desired behaviors. PB space was also separated for both of the goal and the means. The results that the primary goal learned first and the secondary goal learned later is similar to the findings of developmental studies that younger infants tended to imitate only primary goals of actions as ignoring secondary goals.

References

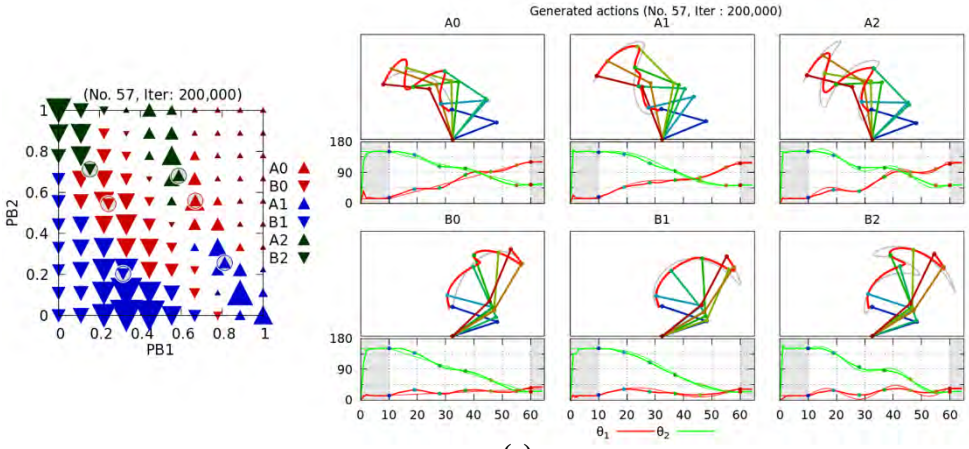
1. Carpenter, M., Akhtar, N., & Tomasello, M. (1998). Fourteen-through 18-month-old infants differentially imitate intentional and accidental actions. Infant Behavior and Development, 21(2), 315-330.
2. Tani, J., Ito, M., Sugita, Y.: Self-organization of distributedly represented multiple behavior schemata in a mirror system: reviews of robot experiments using RNNPB. Neural Networks 17, 1273{1289 (2004).
3. Werbos, P. J. (1990). Backpropagation through time: what it does and how to do it. Proceedings of the IEEE, 78(10), 1550-1560.



(a)



(b)



(c)

Figure 15. Dynamics of PB space and results of action generation. The left side of the figure illustrates which reference actions (from A_0 to B_2) have a minimum error in the PB space. The direction and the color of the triangular markers indicate the goal and the style of movement, respectively. The size of the markers is inversely proportional to the amount of error E_{means}: The larger a marker is, the smaller the error is. Recognized PB values x_{recog} are illustrated as circles with triangular markers inside. The right side of the figure represents the actions generated by the agent and its joint angles. The figures of joint angles represent Y_{ref} (thin lines) for all reference actions Y_{gen} (thick lines) in the time domain. The red and green lines are the first and second joint angles, respectively.

List of Publications and Significant Collaborations

Conference presentations without papers

1. Jun-Cheol Park, Yunhun Jang, Changmo Nam, Jaeyoung Jun, Hyungwon Choi and **Dae-Shik Kim** "Learning an internal representation of a deep convolutional neural network model for pose-based human action recognition", Society for Neuroscience Annual Meeting 2015, Chicago, IL, USA, October 17-21, 2015 (only abstract)
2. Dongpyo Lee, Seungkyu Nam, Hyelin Lim, Sun Mi Park, and **Dae-Shik Kim**, "Decoding sequence of actions using fMRI", Society for Neuroscience Annual Meeting, San Diego, CA, USA, Nov 9-13 2013 (only abstract)
3. Hansol Choi, **Dae-Shik Kim**, "Planning as inference in a Hierarchical Predictive Memory", Proceedings of International Conference on Neural Information Processing, Daegu, South Korea, Nov 3-7 2013 (only abstract)
4. Juhyeon Lee, Jae Hyun Lim, Hyungwon Choi, **Dae-Shik Kim**, "Multiple Kernel Learning with Hierarchical Feature Representations", Proceedings of International Conference on Neural Information Processing, Daegu, South Korea, Nov 3-7 2013 (only abstract)
5. 이주현, 임재현, **김대식**, "Multiple Kernel Deep Network", 대한뇌공학회 2013 뇌와 인공지능 심포지엄, 강원도 하이원리조트 컨벤션호텔, 2013.02.20~21 (only abstract)
6. 박준철, 전재영, 장윤훈, **김대식**, "A Memory Based Learning Method for Chunking and Concatenating Primitive Motor Behaviors", 대한뇌공학회 2013 뇌와 인공지능 심포지엄, 강원도 하이원리조트 컨벤션호텔, 2013.02.20~21 (only abstract)

Manuscripts submitted but not yet published

1. Jun-Cheol Park, **Dae-Shik Kim** and Yukie Nagai, “Learning for goal-directed actions using RNNPB: Developmental change of ‘what to imitate’ ”, Autonomous Robots (submitted)