2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
Congress Center Hamburg
Sept 28 - Oct 2, 2015. Hamburg, Germany

# Neural Network based Model for Visual-motor Integration Learning of Robot's Drawing Behavior: Association of a Drawing Motion from a Drawn Image

Kazuma Sasaki[1], Hadi Tjandra[2], Kuniaki Noda[1], Kuniyuki Takahashi[3], and Tetsuya Ogata[1]

***Abstract*— In this study, we propose a neural network based model for learning a robot's drawing sequences in an unsupervised manner. We focus on the ability to learn visual-motor relationships, which can work as a reusable memory in association of drawing motion from a picture image. Assuming that a humanoid robot can draw a shape on a pen tablet, the proposed model learns drawing sequences, which comprises drawing motion and drawn picture image frames. To learn raw pixel data without any given specific features, we utilized a deep neural network for compressing large dimensional picture images and a continuous time recurrent neural network for integration of motion and picture images. To confirm the ability of the proposed model, we performed an experiment for learning 15 sequences comprising three types of shapes. The model successfully learns all the sequences and can associate a drawing motion from a not trained picture image and a trained picture with similar success. We also show that the proposed model self-organizes its behavior according to types shapes.**

## I. BACKGROUND

Drawing behavior is a core human ability that represents feelings, situations, and knowledge through simple lines. Although drawing activities require complex and diverse cognitive skills to recognize objects or scenes and generate drawing motions, psychological studies have demonstrated that the relationship between vision inputs and motor commands works as the fundamental component in drawing activities [1].

Cognitive developmental robotics [2] has proposed to develop a corresponding computational model for agents to understand such a complex embodiment of cognitive skills. In this study, we focus on a learning model for acquiring visual motor memory through drawing experiences.

Numerous recent studies have demonstrated robots which can draw, paint or write [3][4][5][6]. Mohan et al. proposed a learning model based on catastrophe theory that uses primitive features of shapes called "Critical Points." Their model can decompose when it learns to draw a sequence, and it also generates a drawing motion by synthesizing the primitive features [7]. Mochizuki et al. presented a neuro-dynamical model for acquiring drawing skills of simple shapes through the incremental learning with human robot interactions [8]. They focus on the developmental process of a child's drawing skill. In their study, a robot develops drawing skills from random drawing movements for drawing simple shapes.

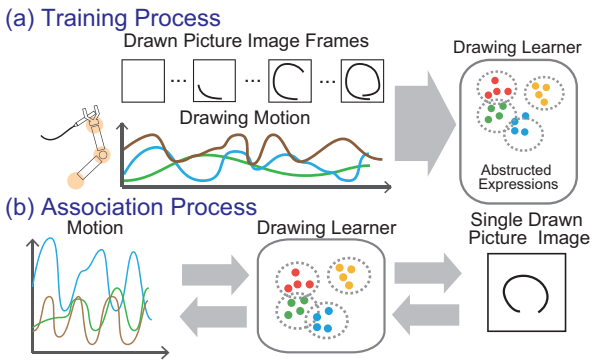


Fig. 1. The overview of the objective system for learning robot's drawing behavior.

The existing approach to learn drawing behavior is to consider a single completed picture as the visual information. However, in the case of humans, the temporal transition of drawn picture image and corresponding motion are reused for drawing and recognition. Babcock et al. suggested that humans use not only visual characteristics of shape but also information about how the picture is produced [9]. This suggestion leads a study about motor-perception of a drawing using visual-motor scheme [10]. In other words, a drawing behavior should include not only drawing motion but also temporal visual features. The visuo-motor experiences of drawing are abstracted as a memory, which allows an agent to associate drawing motion from a picture. In recent time, this types of memory termed "visual motor memory" [11] is starting to be investigated in psychological fields.

The problem to learn temporal visual features is the large calculation cost for every step of raw pixel data. Although Mochizuki et al.'s model considers time series drawn picture's features, their model can learn only the axis of the pen for the picture's features. Therefore, we introduce a learning model which can (1) learn visual sequences as raw image data and (2) associate drawing motion from a picture image using learnt drawing experiences. Fig. 1 shows the overview of the proposed model. The proposed model comprises the training process and association process. In the training process, the drawn picture image and robot's motion are acquired in each step and integrated into visual motor

[1]Kazuma Sasaki, Kuniaki Noda and Tetsuya Ogata are with Graduate School of Fundamental Science and Engineering, Waseda University, Tokyo, Japan ssk.sasaki@suou.waseda.jp

[2]Hadi Tjandra is with Graduate School of Creative Science and Engineering, Waseda University, Tokyo, Japan

[3] Kuniyuki Takahashi is Graduate School of Creative Science and Engineering, Waseda University, Tokyo, Japan, and Japan and Research Fellow of Japan Society for the Promotion of Science (JSPS)

memory as an abstracted expression. These expressions are utilized in association of drawing motion from a target picture image. In addition, those processes are obtained by avoiding elaborating specific features of pictures or motion, and learnt by unsupervised manner.

In this study, we introduce a preliminary learning model on the condition that a robot unicursally draw a shape. To obtain the model, we utilize two artificial neural networks as follows: (1) deep neural network (DNN) and (2) continuous time recurrent neural network (CTRNN). The DNN works as a dimensional compressor for raw image data and faces problems of calculation because of its vastness when optimizing recurrent neural networks. The CTRNN integrates picture image frames and motion by learning both, in sequence. Both networks are trained without shape features or target signals to represent a class of the drawn pictures.

This paper is organized as follows. In Section II, we describe the proposed model. In Section III, we present the practical construction of the proposed model and the drawing experiment for the evaluation. In Section V, we analyze the results and finally we summarize our works in Section V.

## II. ARCHITECTURE

Fig. 2 depicts the proposed model and comprised two neural network models, a Deep Neural Network Autoencoder (DNN Autoencoder) and a Multiple Timescale Recurrent Neural Network (MTRNN). To design the proposed model as a preliminary study, we focus on learning drawing sequences for unicursal shapes.

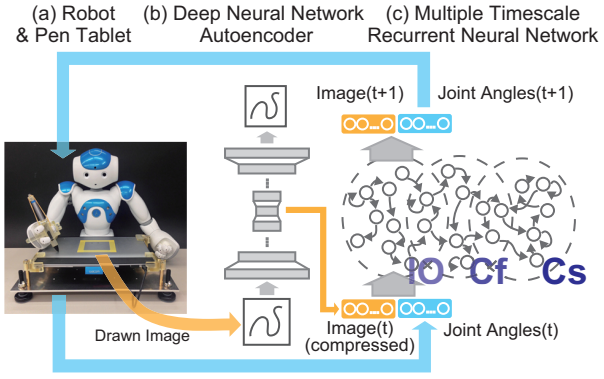


Fig. 2. The proposed model for learning temporal visual-motor sequences which comprises drawing motion and drawn image frames. (a) Robot and Pen Tablet: A robot and pen tablet are fixed to each other via a base plate. (b) Deep Neural Network Autoencoder: A deep neural network model for reducing the dimensions of drawn picture images by learning to encode input data as output. (c) Multiple Timescale Recurrent Neural Network: A continuous recurrent neural networks for learning drawn picture image frames and joint angle values at each frame.

### A. DNN Autoencoder

A deep neural network (DNN) is a feedforward neural network model that has many hidden layers, proposed by Hinton et al [12]. In the proposed model, we apply DNN as an auto-encoder which can learn to encode the input data in the output data. A high-dimensional input layer is fully connected to a small-dimensional central hidden layer through multiple middle hidden layers, and the central hidden layer is reconstructed as the output layer which has the same dimension as the input layer. By training for encoding a set of input data, dimensionally compressed expressions of input data are obtained from the central hidden layer's output. The n-th middle layer's output $\boldsymbol{\epsilon}_n$ is calculated as follows:

$$\boldsymbol{\epsilon}_n = sigmoid(W_{n-1}\boldsymbol{\epsilon}_{n-1} + \boldsymbol{\beta}_{n-1}),$$

where $W$ is the weight matrix and $\boldsymbol{\beta}$ is the bias vector.

The DNN Autoencoder is trained with Hessian-free optimization, which is a truncated-Newton optimization method, proposed by Martens [13]. In the standard Newton method, update amounts of the network's parameters $p_n$ in the n-th update, are found for minimizing the quadratic equation of the cost function $f$:

$$M_{\theta_n}(\theta) = f(\theta_n) + \nabla f(\theta_n)^{\mathrm{T}} p_n + \frac{1}{2} p_n^{\mathrm{T}} B_{\theta_n} p_n,$$

where $\nabla f$ is the gradient of $f$ and $B$ is a damped Hessian matrix of $f$. In the Hessian-free approach, a positive semi-definite Gauss-Newton curvature matrix obtained in the linear conjugate gradient for $M_{\theta_n}(\theta)$ is used instead of matrix $B$ which is extremely expensive when network size becomes large.

In the proposed model, we applied DNN Autoencoder for reducing the dimensions of drawn picture frames into small-dimensional image features. In the training process, all drawn picture frames are trained by the DNN Autoencoder. After this training, we prepare training data sequences which have dimensionally compressed image frames and time series of joint angles.

### B. MTRNN

The multiple timescales recurrent neural network (MTRNN) is a continuous time recurrent neural network model (CTRNN) [14]. The CTRNN is known as a dynamical systems model of neural networks. Activities of the CTRNN's neurons are decided by not only synaptic inputs but also by the past history of the neural state. The firing rate of neuron $\dot{u}_{i,t}$, which has the time constant $\tau_i$ is described as follows:

$$\tau_i \dot{u}_{i,t} = -u_{i,t} + \sum_j w_{ij} x_{j,t},$$

where $u_{i,t}$ is internal state of $i$th neuron in $t$th step and $w_{ij}$ is the weight from activation of $j$th neuron $x_{j,t}$ to $i$th neuron. Neurons of the CTRNN are divided into input-output neurons (IO unit) and non-input-output neurons, and termed context units. In the proposed model, the CTRNN predicts the next step of input data as output data in IO unit. The context unit influences the activities of an IO unit through weights between them.

In the case of the MTRNN, it has more than two types of context units comprising the fast context units and the slow context units that have different time constant values.

Because of the difference between these values, the dynamics of trained sequences are effectively memorized as the combination of fast changing dynamics in the fast context units which have a smaller constant value, and slow changing dynamics in the slow context units which have a larger constant value (Fig. 3).

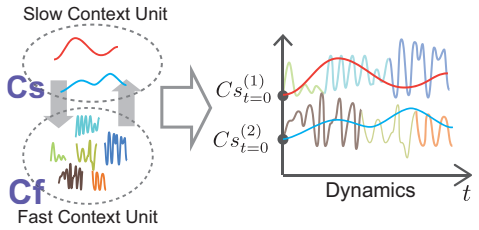


Fig. 3. The role of context unit and initial value of $Cs$.

MTRNN is trained using a back propagation through time algorithm (BPTT), which is a general optimization method for recurrent neural networks [15]. A parameter of the MTRNN in $n$th iteration $\theta_n$ is decided for minimizing the mean square error $E(\theta_n)$ with parameters of network $\theta_n$ as shown in the following mathematical expression.

$$\theta_{n+1} = \theta_n - \alpha \frac{\partial E(\theta_n)}{\partial \theta_n},$$

Using BPTT, MTRNN can learn a number of sequences of dynamical behavior and self-organize them through the recurrent process. When MTRNN predicts a sequence, the initial activation value of neurons in context units decide the whole direction of the predicted sequence. Subsequently, we call this initial value the "initial context value." However, recurrent neural network models including MTRNN have a limitation for learning large dimensional data like an image's raw pixel data because of large calculation cost in the recurrent process. Therefore, we use dimensionally compressed drawn picture image frames from a DNN Autoencoder as raw pixel image data.

In the training process, MTRNN learns visual-motor sequences which comprise dimensionally compressed drawn picture image frames and time series of joint angles. For ease of the association process, $Cf_{t=0}$ (Fast context unit) is fixed with zero in the training process. Therefore, learning parameters $\theta$ represents the weight for each neuron $w$ and initial context values $Cs_i$ that corresponds to each trained sequence.

### C. Association Process

The association process of the proposed model involves searching for an appropriate initial context value of slow context units $Cs_{t=0}$ as depicted in Fig.4. First, a white picture image and a target picture image are dimensionally compressed by the trained DNN Autoencoder. Next, the trained MTRNN predicts and generates a sequence which starts from the white picture image's features with a tentative initial slow context value $Cs_{t=0}$ for the given arbitrary steps. In this generation, the initial joint angles come from initial position of the given right arm of the robot. After generating a temporary sequence, $Cs_{t=0}$ is re-trained using BPTT to minimize error between the target picture image features and the generated drawn picture image by MTRNN. Therefore, the re-trained initial context value becomes an appropriate value for drawing a not trained target picture image under the conditions of a given initial position and step length.

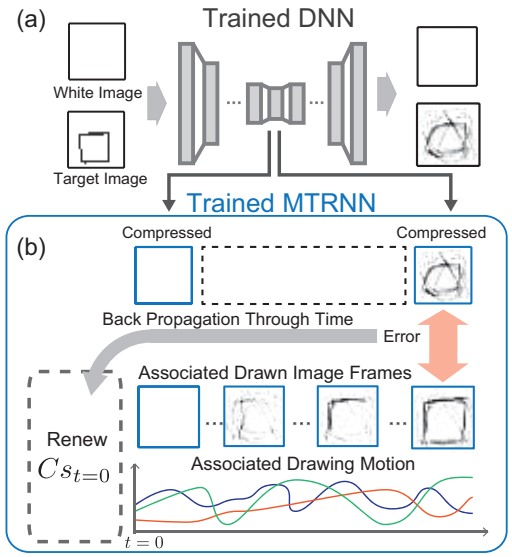


Fig. 4. The method of associating motion for a target drawn image. (a) The target image and a white image are compressed by trained DNN. (b) Searching for the initial value of the slow context unit with minimal error between the target image and generated drawn image frame in the final step of the trained MTRNN's forward propagation from the compressed white image.

## III. EXPERIMENT

The experiment is designed to confirm the ability of the proposed model by a simple drawing task with a small humanoid robot, NAO, developed by Aldebaran Robotics [16]. An intuos-pen tablet [17] is used for capturing drawn picture images. The tablet and the pen are fixed to a base plate and the right hand, respectively (Fig. 5). To avoid capturing errors imparted by the pen tip lifting from the tablet, the pen is allowed to move vertically.

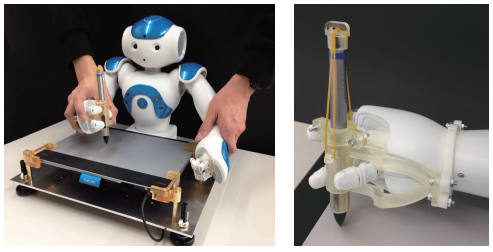

Fig. 5. Experiment setup. A humanoid robot fixed with a pen tablet through a plate.

For the training process, we record 15 sequences of data of drawn picture image frames and joint angles during direct teaching. These drawing sequences comprise three types of

TABLE I

NUMBER OF TRAINING PICTURE IMAGES AND EXPERIMENTAL PARAMETERS.

| | IO | DIMS | DATA(Recorded) | TRANS | ROTATED | Training iterations |
|---|---|---|---|---|---|---|
| DNN Autoencoder | 900 | 400-180-80-30-10-30-80-190-400 | 494 | 31940 | 2910 | 100 |
| MTRNN | 15($\tau = 1$) | Cf ($\tau = 12$): 30, Cs ($\tau = 60$): 20 | 494 | - | - | 15000 |

*The IO, DIMS, DATA, TRANS, ROTATED, and Training Iter give the dimension of IO neurons, the dimensional structure of the networks, the number of the recorded training data, the translated training data, the rotated training data, and the iteration for the optimization, respectively.

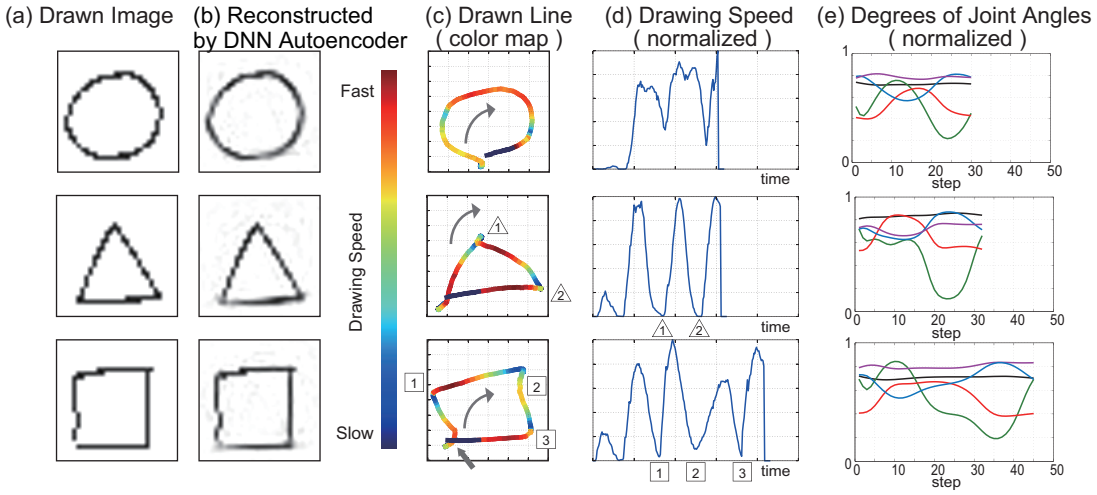


Fig. 6. Summary of the generation results of the training sequences by the trained proposed model. (a) Drawn Image: Final step of drawn picture image frame in the training data sequences. (b) Reconstructed by DNN Autoencoder: Reconstructed the final step of drawn image frame. (c) Drawn Line: Drawn lines by the robot and the lines colored by speed of the pen's position. (d) Drawing Speed: Speed of pen's position with numbers corresponding to the corners of the drawn lines, respectively. (e) Degrees of Joint Angles: Time series joint angles generated by MTRNN. Line colors correspond to each DOF of the robot's right arm.

drawing experiences: circles, triangles, and squares. The five variations for each shape have almost the same beginning position and all pictures are drawn clockwise with one stroke. The length of the recorded drawing sequences is between 25 and 50 steps (equivalent to between five to ten seconds). Triangles and squares have two or three corners, and those strokes appear as dramatical speed reductions. For recording the training sequences, five DoF of the robot's right arm are utilized for time series joint angles, and $30 \times 30$ binary pixels are captured as drawn picture image frames.

After recording training sequences, the proposed model is trained by the recorded training sequences. The training data set for the DNN Autoencoder includes not only all frames of drawn picture images but also rotated and translated versions of these picture frames because the acquired image features has to have the generalization capability of the shape's variation. The network is trained for 100 iterations.

The training data and the structure are summarized in TABLE I. After the training of the DNN Autoencoder, MTRNN is trained using 15 sequences of dimensionally compressed drawn picture image features and joint angles. The structure of the MTRNN is also presented in TABLE I. In this training, the initial value of $Cf$ and the remaining part of $Cs$ are fixed with zero.

Next, MTRNN associates three drawing sequences from not trained picture images as shown in Fig. 7 (a). The length of each association is 45 steps and the initial joint angles are common among all association sequences. The retraining iterations for searching the initial value of $Cs$ is 2000 times for each sequence, and the initial values of 15 dimensions of $Cs$ and $Cf$ are fixed with a zero vector, the as same as in the training process.

Finally, MTRNN generates three sequences with searched initial values of $Cs$ for 45 steps and the robot draws all sequences operated by the generated joint angles respectively.

## IV. RESULTS

### A. Generation of the training sequences

First, the proposed model generates the training sequences to confirm the ability to memorize them. Fig. 6 represents three examples of the drawing motions from generated time series joint angles. These examples represent each of the three shapes. Note that (b) is a reconstruction of drawn image (a) by the DNN Autoencoder. We also depict the drawn lines and pen's speed as evaluated by the distance between each step. Figures in the third column (c) show the drawn line followed by the generated time series of joint angles, colored according to the speed of the pen tip, which is shown in the fourth column (d). The drawn results clearly show that the drawn pictures keep their shapes and the characteristics of drawing motions show speed reduction of the pen tip in the corners marked as numbers in the figures.

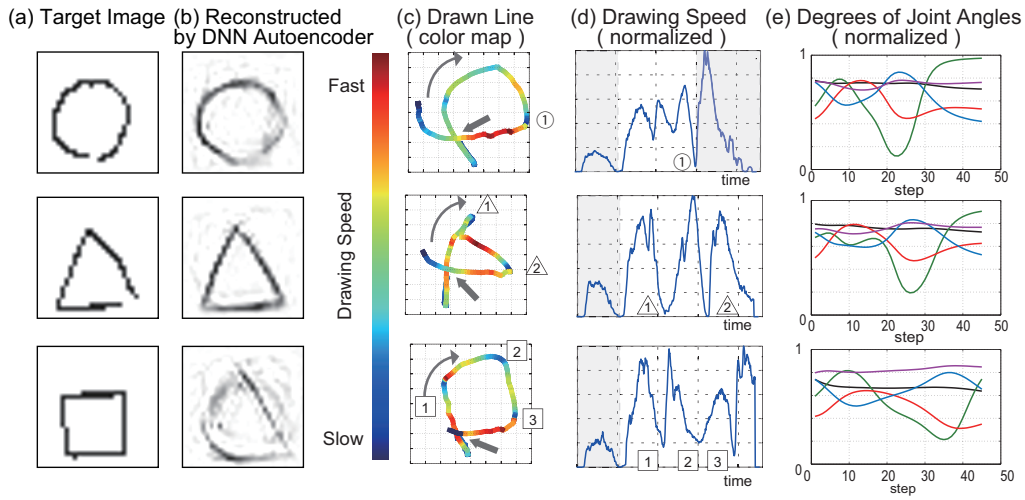


Fig. 7. Associated drawing sequences. (a) Target Image: Images to associate the drawing sequence. (b) Reconstructed by DNN Autoencoder: Reconstructed the final step of drawn image frame. (c) Drawn Line: Drawn lines by the robot and the lines colored by speed of the pen's position. (d) Drawing Speed: Speed of pen's position with numbers corresponding to the corners of the drawn lines. (e) Degrees of Joint Angles: Time series joint angles associated by MTRNN. Line colors correspond to each DOF of the robot's right arm.

### B. Drawing Result by the Associated Motion

Fig. 7 shows the results of associated drawing sequences in the same manner as Fig. 6. In the results of the associated drawn picture image by the trained DNN, circle and triangle clearly remain in their shapes in contrast with results of the square. Furthermore, the drawn pictures by the associated motions include some distortions but they keep their shapes.

The associated picture image of the square was not well reconstructed. We hypothesize that this is because the training data set does not include a scaled version of the original picture images. The DNN model is not learnt the bigger squares than the trained squares. However, the model can successfully adapt its dynamics using a difference of compressed feature in the association process and generated the drawing motion. The main reason of the distortions at the edge, initial, and ending points in the drawing results are also the characteristics of the model. The CTRNN cannot completely recover the trajectory of the arm which changes non-continuously because CTRNN tend to replicate the training data by continuous dynamics.

Although the drawn results include distortions, associated motion includes the common characteristics of chaining speed with the training dataset. In the case of the circle, the pen tip draws the upper half of the circle while changing its speed slightly and stops after around 30 steps. Then it moves again, and the sequence ends on the left side of the picture frame. The pen speeds in the triangle and square are more similar to the generation results of training sequences. The speed slows down clearly at corners and drawn lines depict shapes of a target picture image.

### C. The Features of the Proposed Model

Furthermore, to summarize associated drawing motions, we also analyze the behavior of the proposed model when the model generates the training data set and associations.

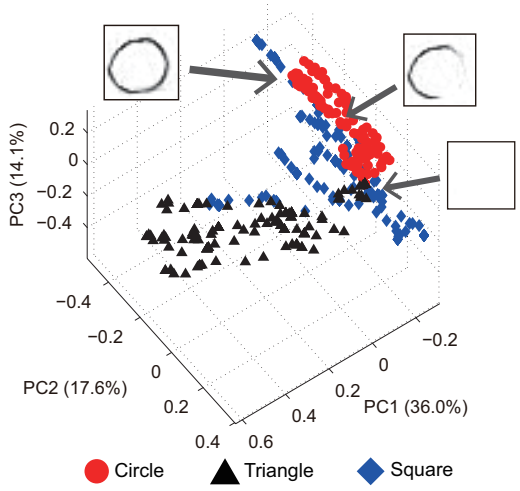


Fig. 8. Acquired features of drawn picture image frames (Training data set). PC1 to PC3 axes correspond to principal components 1–3, respectively.

Fig.8 depicts the acquired dimensionally compressed image features of principal components of the training data set. Acquired features are clearly separated by types of shapes, with features of one sequence forming one line for each the shape which, however start from a common point for all sequences: i.e., the point of the white image feature. As features are separated from the point of the white image, they divide according to shape and approach the end of each sequence.

Fig.9 shows the value of fast context units $Cf$ graphed according to types of shape. The feature's lines are similar, according to types of shape. Just as with the DNN Autoencoder, all lines start from a common point which expresses the beginning of drawing sequence.

Fig.10 shows the value of slow context unit $Cs$ com-

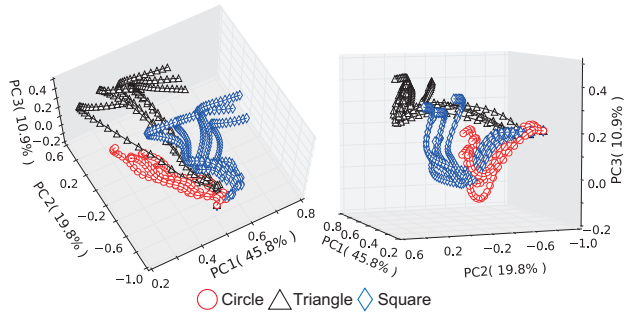


Fig. 9. Acquired features of visual motor experiences in fast context units of the MTRNN in the case of the training sequences. These two figures correspond to the same 3D-plots viewed from different angles. The PC1–PC3 axes correspond to the principal components 1–3 with their respective contribution values.

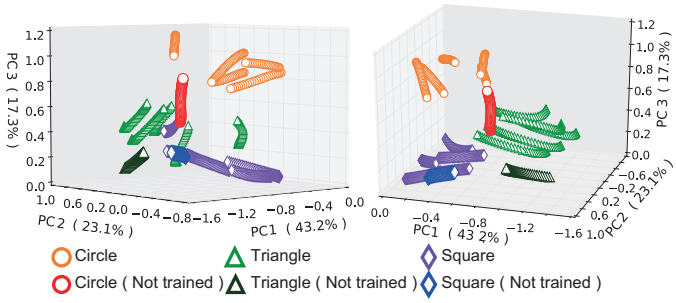


Fig. 10. Acquired features of the visual motor experiences in slow-context units of the MTRNN in the training and the associated sequences. These two figures correspond to the same 3D-plots viewed from different angles. The PC1–PC3 axes correspond to the principal components 1–3 with their respective contribution values.

pressed by principal component analysis. The values clearly differentiated by types of shape and they start at different positions. Associated values are located near the value which has the same shape type. It should be emphasized that the initial values of the sequence in the slow context unit differ from each other, although the initial values of the fast context unit share the same value. This means that the trained drawing behaviors are self-organized by MTRNN, and the slow context unit value represents types of shapes. Therefore, MTRNN can associates types of behavior which correspond to a target image picture.

## V. CONCLUSION

In this study, we propose a neural network based model for learning drawing as a dynamic behavior comprising time series of raw picture image frames and motion. For solving the demanding calculation of raw image data, we apply a deep neural network which works as a dimensional compressor optimized with an unsupervised learning. Instead of learning raw drawing picture images, a continuous time recurrent neural network model, MTRNN learns the dimensionally compressed image features and joint angles.

Through an experiment of learning 15 drawing sequences, of three types of simple shapes, we show that the proposed model successfully learns these sequences and organize them by the type. Learnt sequences are self-organized through the recurrent process which has multi time scales. As shown in the results, the fast context unit represents shape-based features and the slow context unit represents discriminating dynamics of the fast context unit. Further, the proposed model can also associate drawing motion from a drawn picture by changing its dynamics to draw the target picture. One of the challenges for future study is learning drawing sequences which have multi strokes.

## ACKNOWLEDGMENT

The work has been supported by JST PRESTO "Information Environment and Humans" and MEXT Grant-in-Aid for Scientific Research on Innovative Areas "Constructive Developmental Science" (24119003).

## REFERENCES

[1] S. McCrea, "A neuropsychological model of free-drawing from memory in constructional apraxia: A theoretical review," *American Journal of Psychiatry and Neuroscience*, vol. 2, no. 5, pp. 60–75, 2014.

[2] M. Asada, K. F. MacDorman, H. Ishiguro, and Y. Kuniyoshi, "Cognitive developmental robotics as a new paradigm for the design of humanoid robots," *Robotics and Autonomous Systems*, vol. 37, no. 2–3, pp. 185–193, 2001.

[3] G. Jean-Pierrem and Z. Said, "The artist robot: a robot drawing like a human artist," in *Proc. IEEE International Conference on Industrial Technology (ICIT'12)*, Athens, Greek, Mar. 2012, pp. 486–491.

[4] S. Mueller, N. Huebel, M. Waibel, and R. D'Andrea, "Robotic calligraphy – learning how to write single strokes of chinese and japanese characters," in *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'13)*, Tokyo, Japan, Nov. 2013, pp. 1734–1739.

[5] S. Calinon, J. Epiney, and A. Billard, "A humanoid robot drawing human portraits," in *Proc. IEEE International Conference on Humanoid Robots (Humanoids'05)*, Tsukuba, Japan, Dec. 2005, pp. 161–166.

[6] S. Kudoh, K. Ogawara, M. Ruchanurucksc, and K. Ikeuchi, "Painting robot with multi-fingered hands and stereo vision," *Robotics and Autonomous Systems*, vol. 57, no. 3, pp. 279–288, 2009.

[7] V. Mohan, P. Morasso, J. Zenzeri, G. Metta, V. S. Chakravarthy, and G. Sandini, "Teaching a humanoid robot to draw 'shapes'," *Autonomous Robots*, vol. 31, no. 1, pp. 21–53, 2011.

[8] K. Mochizuki, S. Nishide, H. Okuno, and T. Ogata, "Developmental human-robot imitation learning of drawing with a neuro dynamical system," in *Proc. IEEE International Conference on Systems, Man, and Cybernetics (SMC'13)*, Manchester, England, Oct. 2013, pp. 2336–2341.

[9] M.Babcock and J. Freyd, "Perception of dynamic information in static hand-written forms," *The American journal of psychology*, vol. 101, no. 1, pp. 111–130, 1988.

[10] A. Pignocchi, "How the intentions of the draftsman shape perception of a drawing," *Consciousness and cognition*, vol. 19, no. 4, pp. 887–898, 2010.

[11] A. H. Waterman, J. Havelka, P. R. Culmer, L. J. B. Hill, and M. Mon-Williams, "The ontogeny of visual – motor memory and its importance in handwriting and reading : a developing construct," *Proceedings B of Biological Science in Royal Society*, vol. 282, 2015.

[12] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, pp. 504–507, 2006.

[13] J. Martens, "Deep learning via hessian-free optimization," in *Proc. 27th International Conference on Machine Learning (ICML'10)*, Haifa, Israel, June 2010, pp. 735–742.

[14] Y. Yamashita and J. Tani, "Emergence of functional hierarchy in a multiple timescale neural network model: a humanoid robot experiment," *PLOS Computational Biology*, vol. 4, no. 11, 2008.

[15] P. J. Werbos, "Backpropagation through time: what it does and how to do it," *Proceedings of the IEEE*, vol. 78, no. 10, pp. 1550–1560, 1990.

[16] (2015, Feb.) Aldebaran robotics nao data sheet. [Online]. Available: https://www.aldebaran.com/en/solutions/documentation

[17] (2015, Feb.) Wacom intuos pen & touch small. [Online]. Available: http://www.wacom.com/en-us/products/pen-tablets/intuos-pen