Development of Functional Hierarchies for Visual Recognition and Action Generation: "Synergetic" Coordination of them in Humanoid Robots

Jun Tani

KAIST

http://neurorobot.kaist.ac.kr/

Self-organization of functional structures in cognitive development of humanoid robots

- It is hard to program all complex motor patterns of humanoids with high DOF.
- It is hard to achieve semantic level understanding of visual streams in pixel level by programming.
- Recent success in deep learning suggests that all necessary functional structures for humanoids could be developed via iterative learning of own multimodal perceptual experience.

Today’s Topics

- Learning to generate compositional actions by humanoid robots (short review).
  - Development of temporal hierarchy for action
- Learning to recognize dynamic visual image of human actions (more focus).
  - Development of spatio-temporal hierarchy for vision.
- Development of “Synergy” in Visuo-Motor-Attentional coordination by humanoid robots.
  - Synergetic integration of the aforementioned two models.

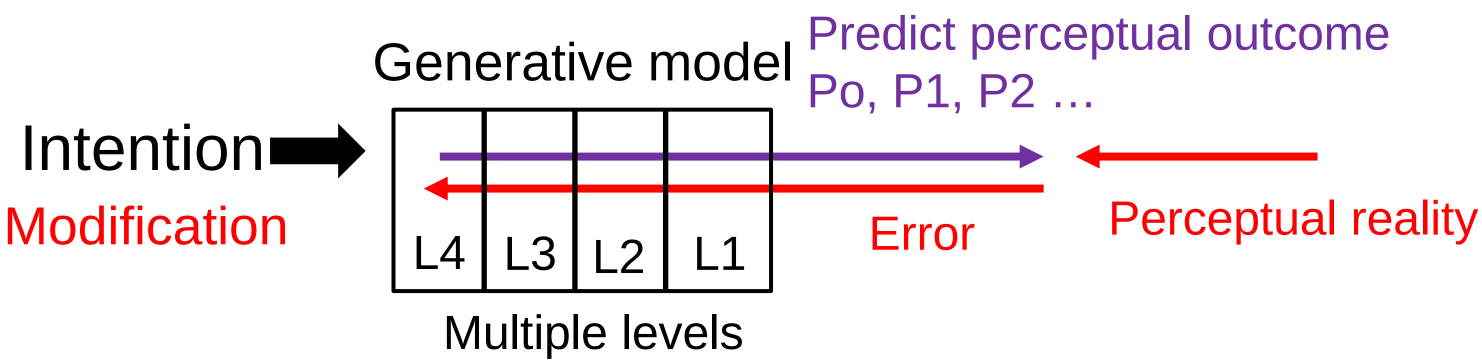
Hypothesis

- If adequate spatio-temporal constraints are imposed on dynamic activity in neural networks, necessary functional hierarchy might be developed in the course of consolidative learning of experiences.
- The essential mechanisms might be well accounted by dynamical sys. language.
  - Different classes of attractors
  - Parameter bifurcation
  - Initial sensitivity

Development of functional hierarchy for action generation

Predictive-Coding for Learning, Generation and Recognition

(Rao & Ballad, 1999; Tani, 1999, 2003, 2014; Friston, 2007)



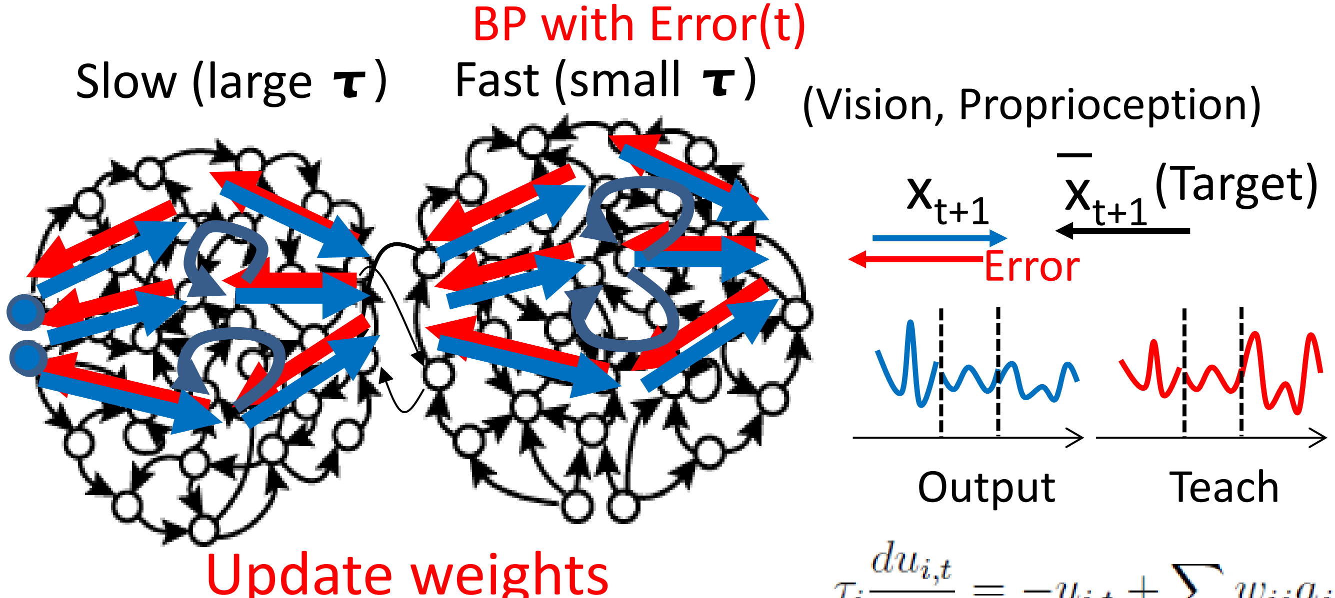
Change connectivity weights and intention in the direction of minimizing prediction error!!

Self-Organization of Functional Hierarchy in Multiple Timescales RNN (MTRNN)

(Tani 2003; Paine & Tani, 2004; Yamashita & Tani, 2008)

(0.2, 0.8)

Update initial states (intentions) in slow net

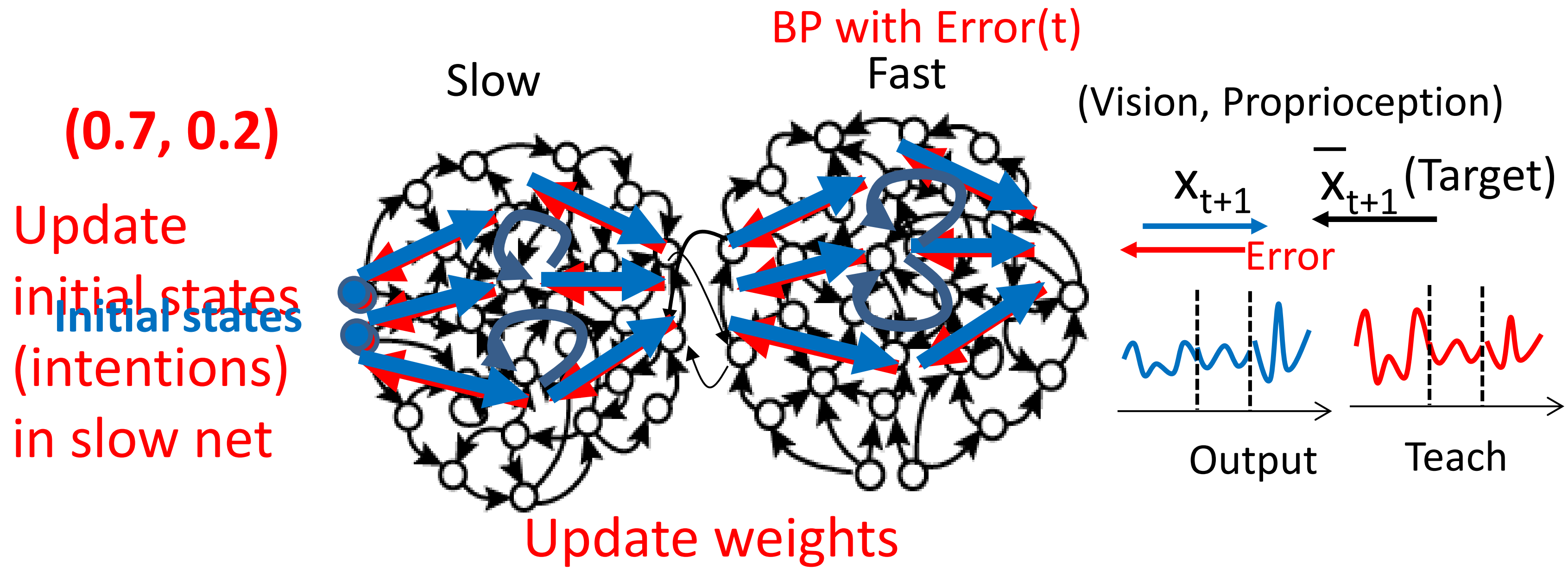


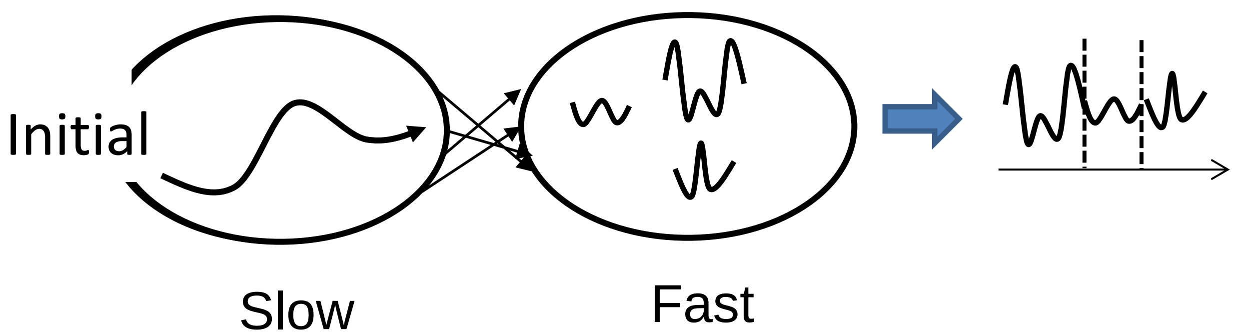
$$\tau_i \frac{du_{i,t}}{dt} = -u_{i,t} + \sum_j w_{ij} a_{j,t}$$

$$a_i = sigmoid(u_i)$$

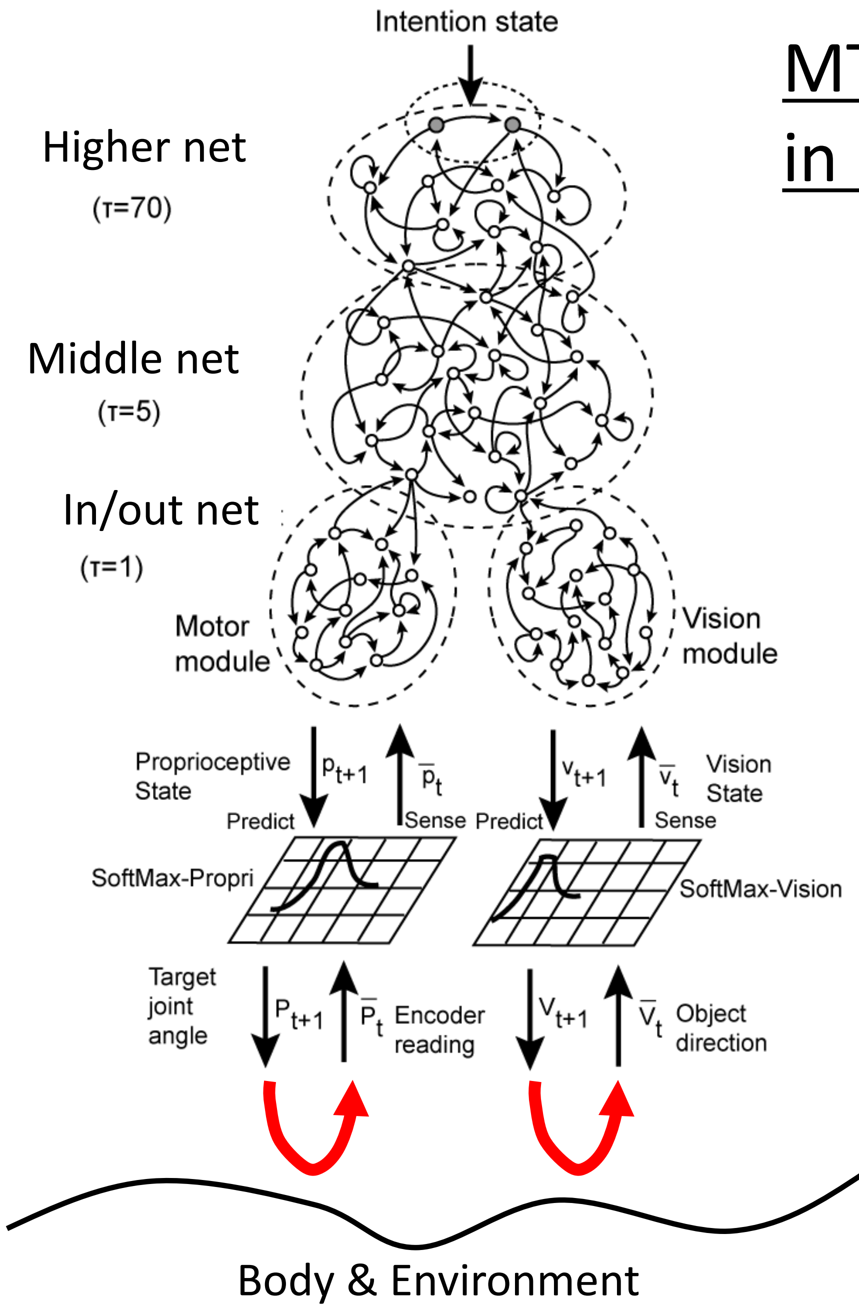
Self-Organization of Functional Hierarchy in Multiple Timescales RNN (MTRNN)

(Paine & Tani, 2004; Yamashita & Tani, 2008)





MTRNN architecture used in robotics experiments



Closed-loop for mental simulation

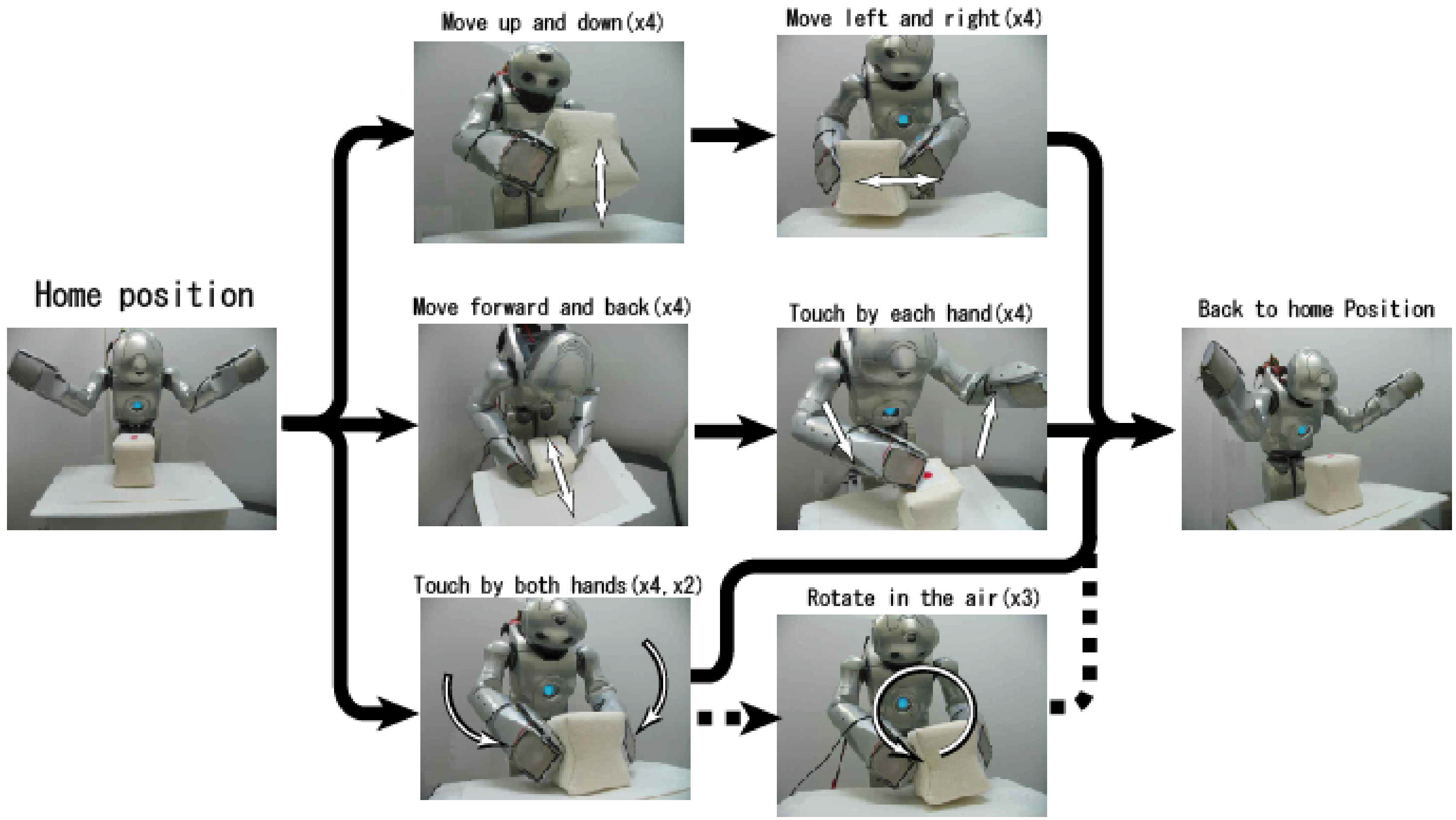
Three Different Goal-Directed Tasks Are Simultaneously Trained

(Nishimoto & Tani 2009)

Task 1

Task 2

Task 3



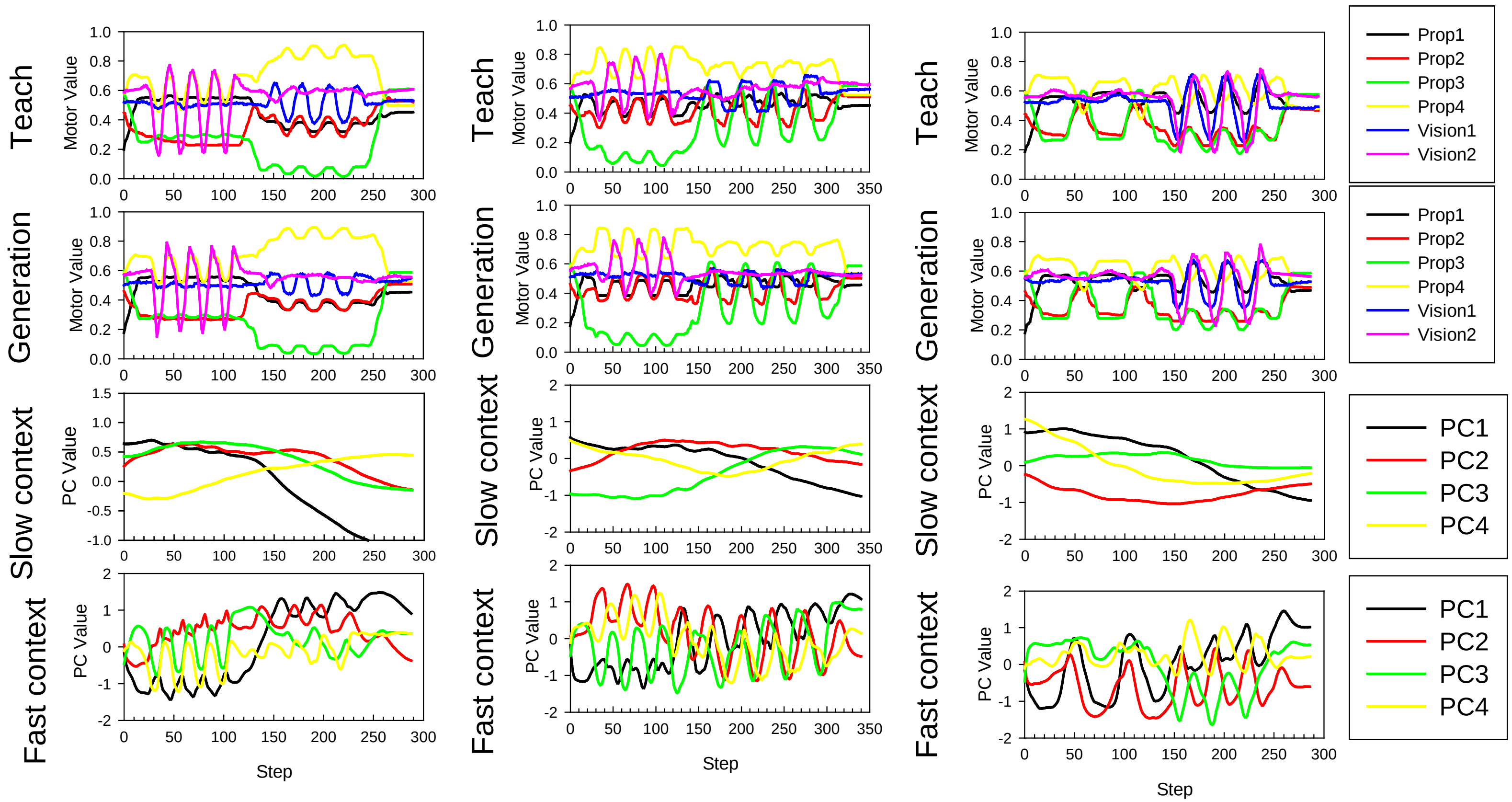
Interactive Tutoring Video

Test generation after learning

After the 3rd tutoring

(One more)

All 3 task sequences at the end of the final tutoring session

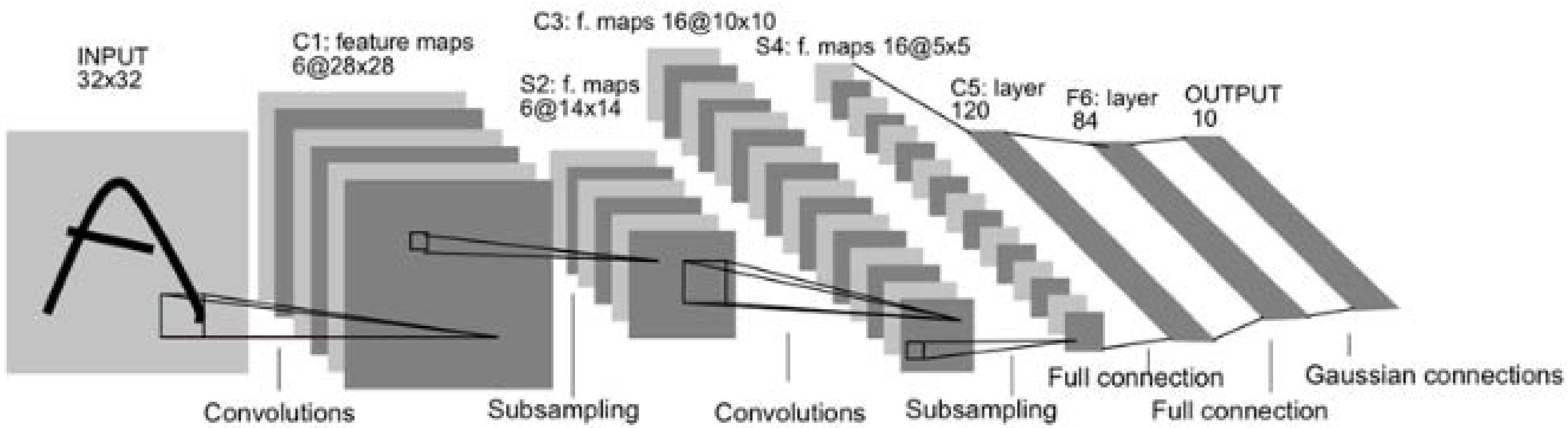


(a) Task 1 (b) Task 2 (c) Task 3

"Kinetic Melody" by Luria

Visual Categorization/Recognition of Human Actions via Learning of Exemplar

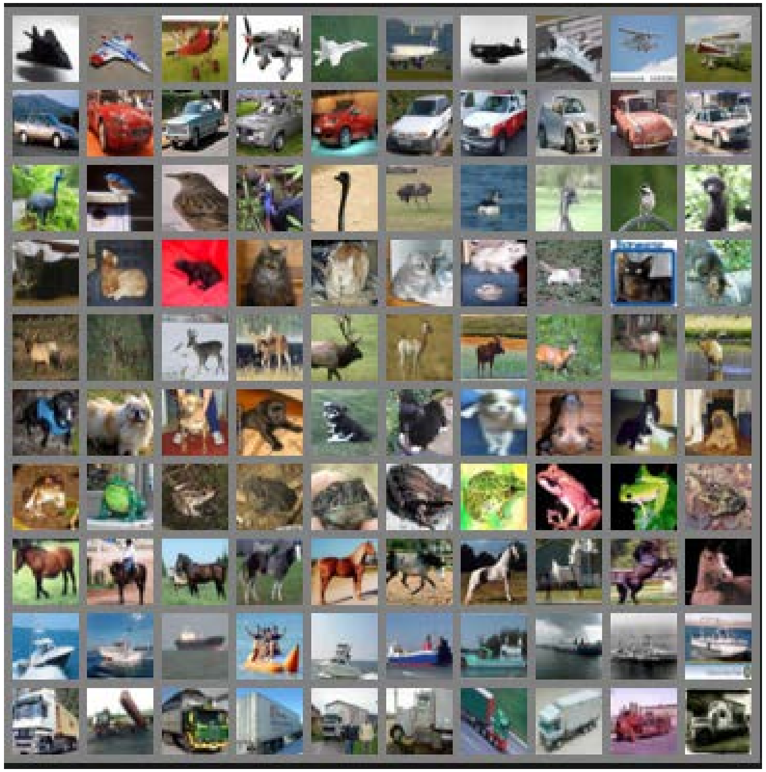
Prior-Study:
Convolution Neural Network (CNN)
for Categorization of Static Visual Patterns

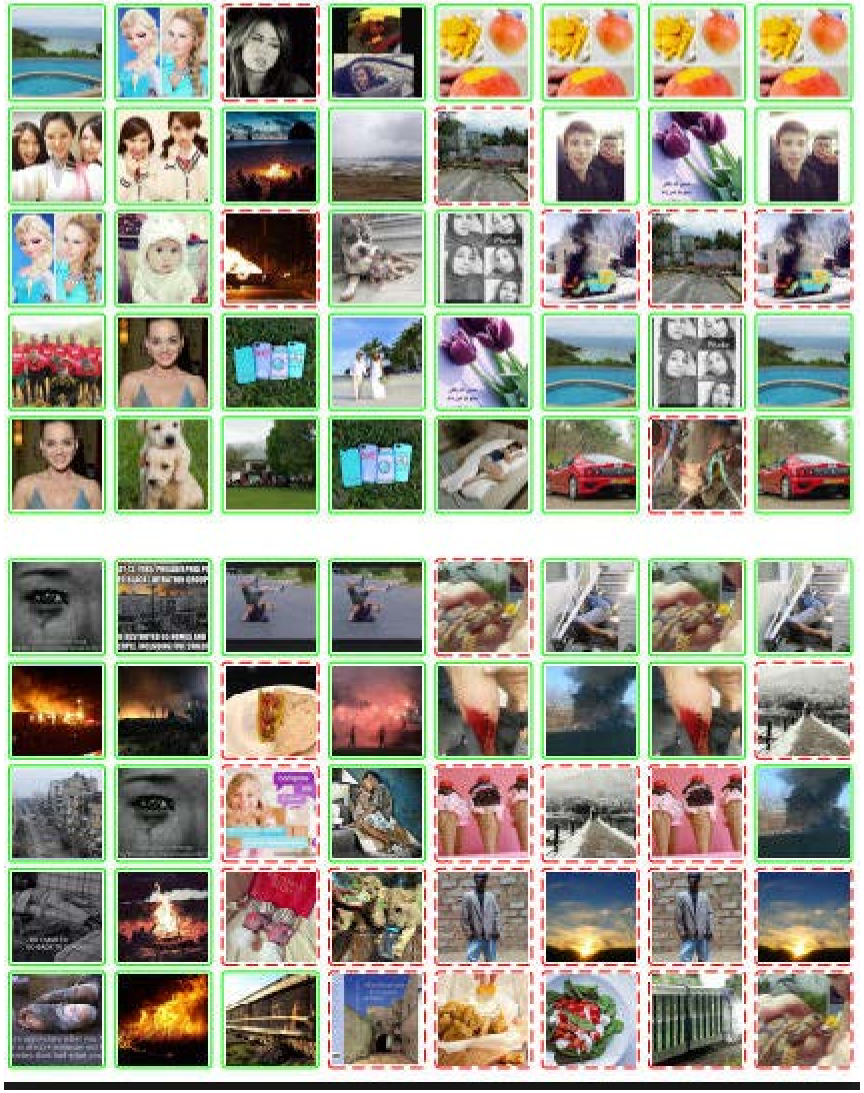


(LeCun et al., 1998)

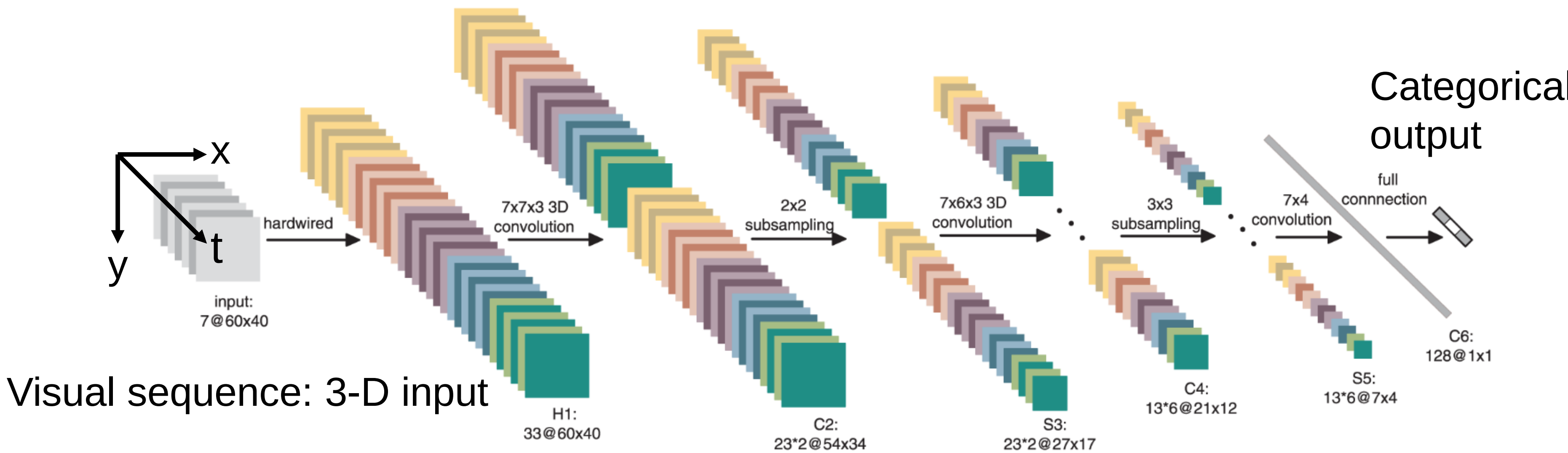
Recent CNN with 30 layers trained with 1 million of visual image in ImageNet can classify hundreds of object image with error rate of 0.0665. The CNN's performance in this task is close to that of human (Wikipedia).

Example training dataset for CNN





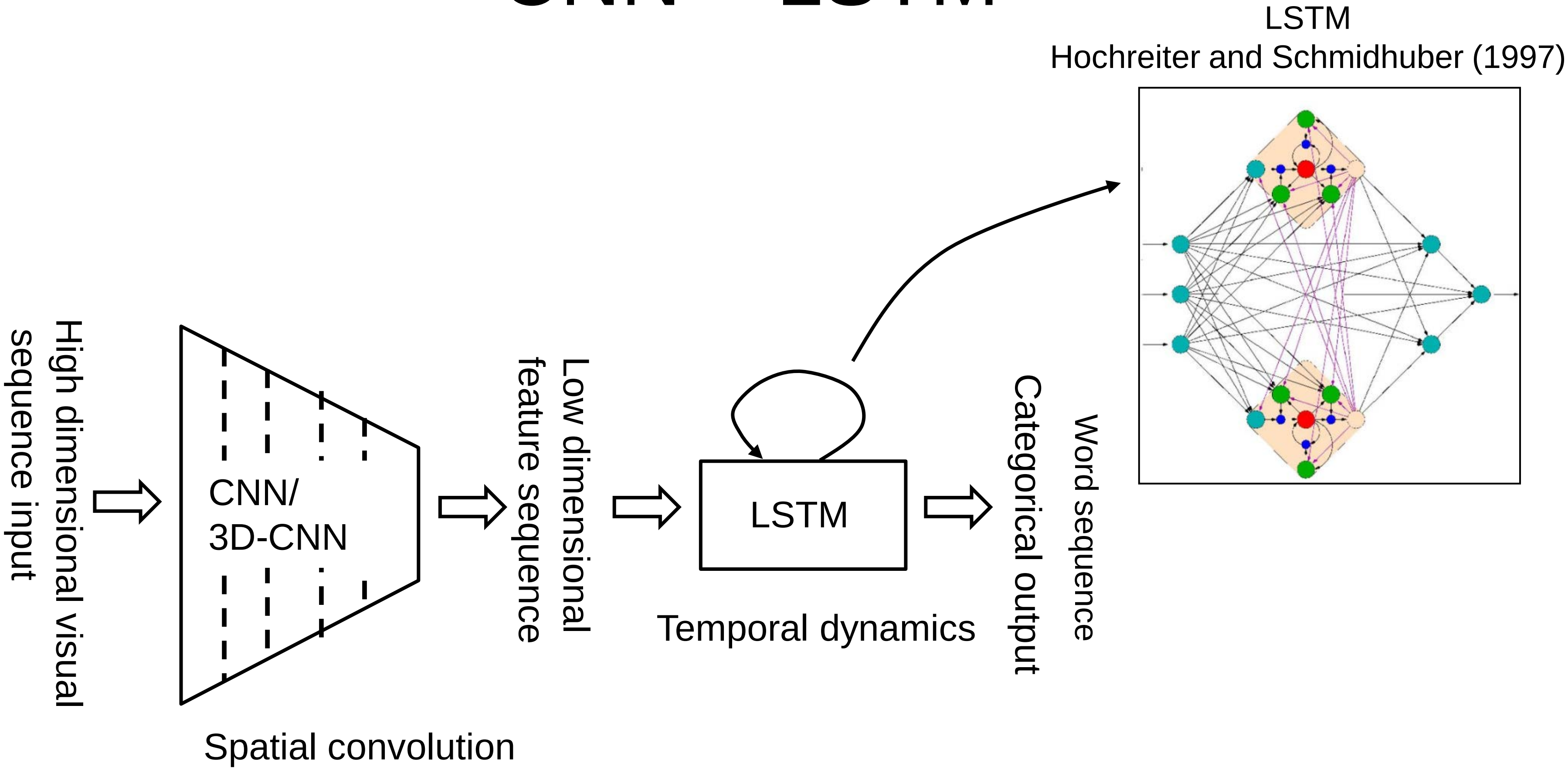
Prior-Study:
3-D CNN for Categorization of Dynamic Visual Patterns



(Baccouche et al., 2011; Karpathy et al., 2014)

Brute force…

CNN + LSTM



(Baccouche et al., 2011; Donahue et al., 2015)

Spatial computation and temporal one are performed separately…

(Donahue et al., 2015)

Language output: “Cat is playing with a toy”

Does this require temporal information?

For achieving semantic-level visual recognition capability, some context-dependent information processing mechanism should be indispensable.

Development of Spatio-Temporal Hierarchy in Learning Dynamic Visual Streams

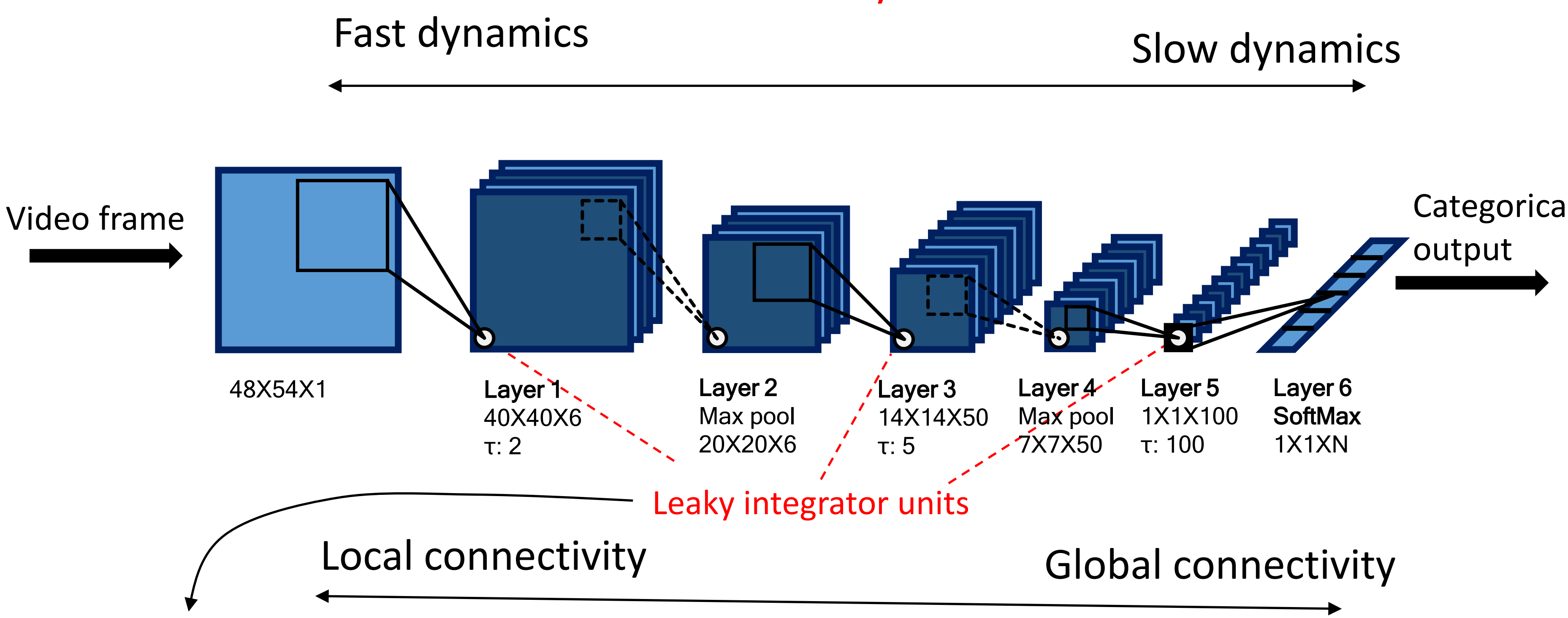
(Jung, Hwang & Tani, 2015)

- Imposing multiple scales spatio-temporal constrains on neural activity
  - ***Spatial*** constraints by ***convolution kernel size*** differences
    - Local connectivity to global one.
  - ***Temporal*** constrains by ***timescale*** differences
    - Fast to slow
- Self-organization of spatio-temporal hierarchy
- Visual recognition of compositional human actions.

Multiple Spatio-Temporal Scales NN (MSTNN) for Dynamic Vision

(Jung, Hwan, Tani, 2015)

A Dorsal Pathway Model

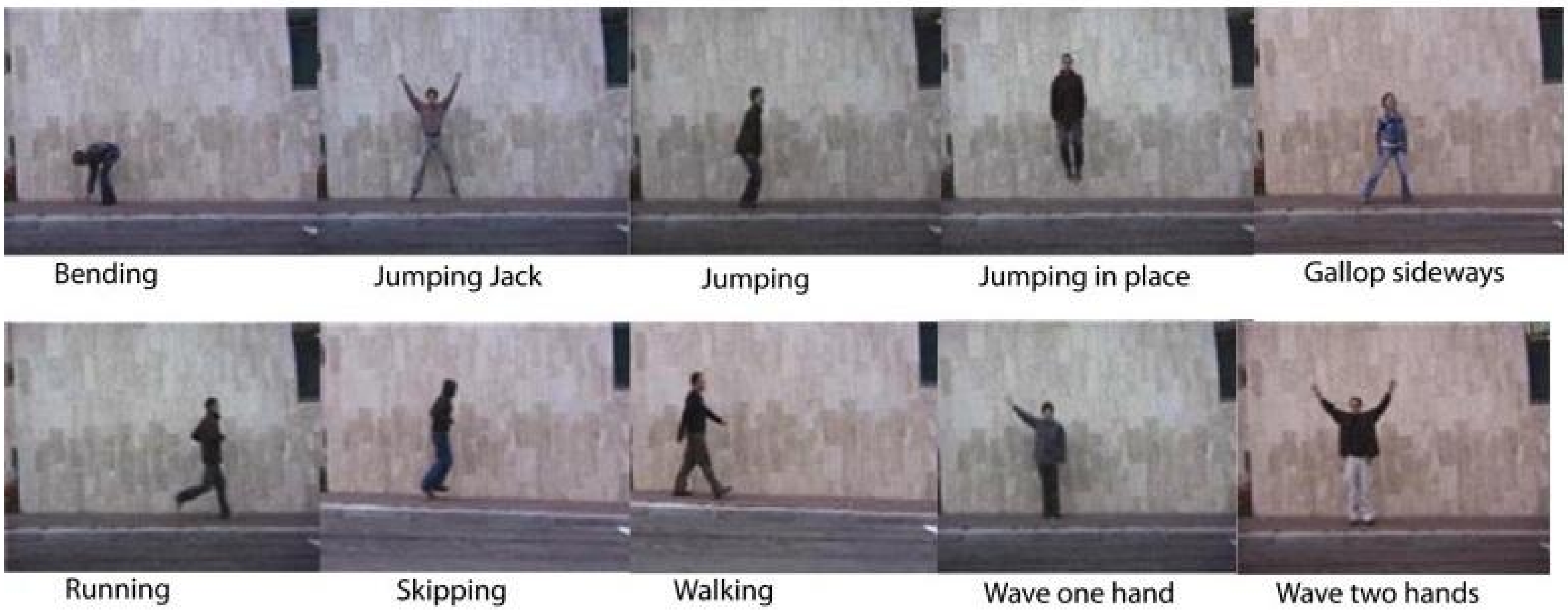


$$\tau_i \frac{du_{i,t}}{dt} = -u_{i,t} + \sum_j w_{ij} a_{j,t}$$

$$a_i = sigmoid(u_i)$$

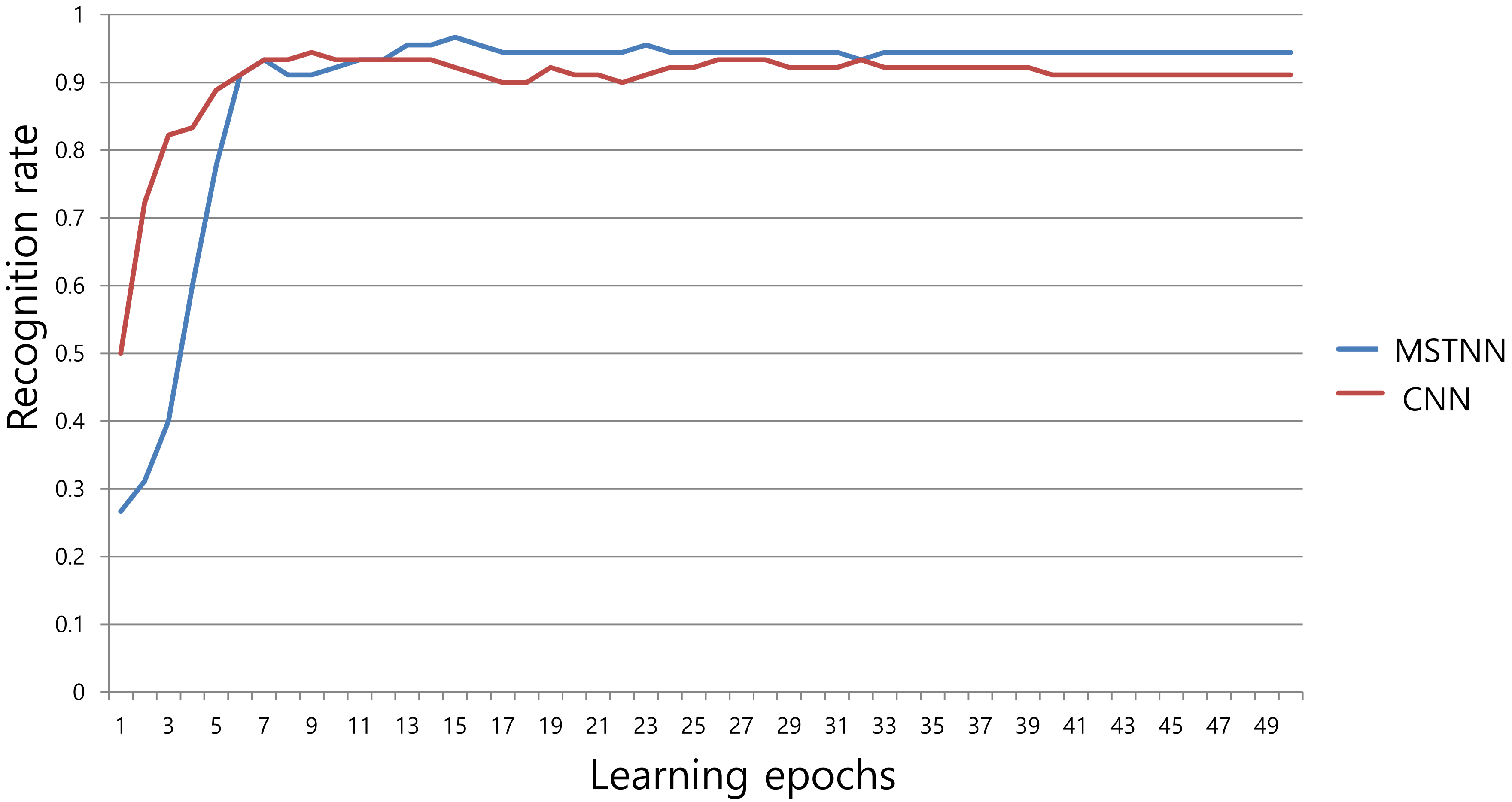
Ideas of CNN (LeCun)

Weizmann Video Dataset of Intransitive Actions



- # of videos 90, # of behaviors 10, # of subjects 9.
- Learning to recognize behavior categories.
- Leave one cross validation (Training with 8 subjects, test with one remained subject).

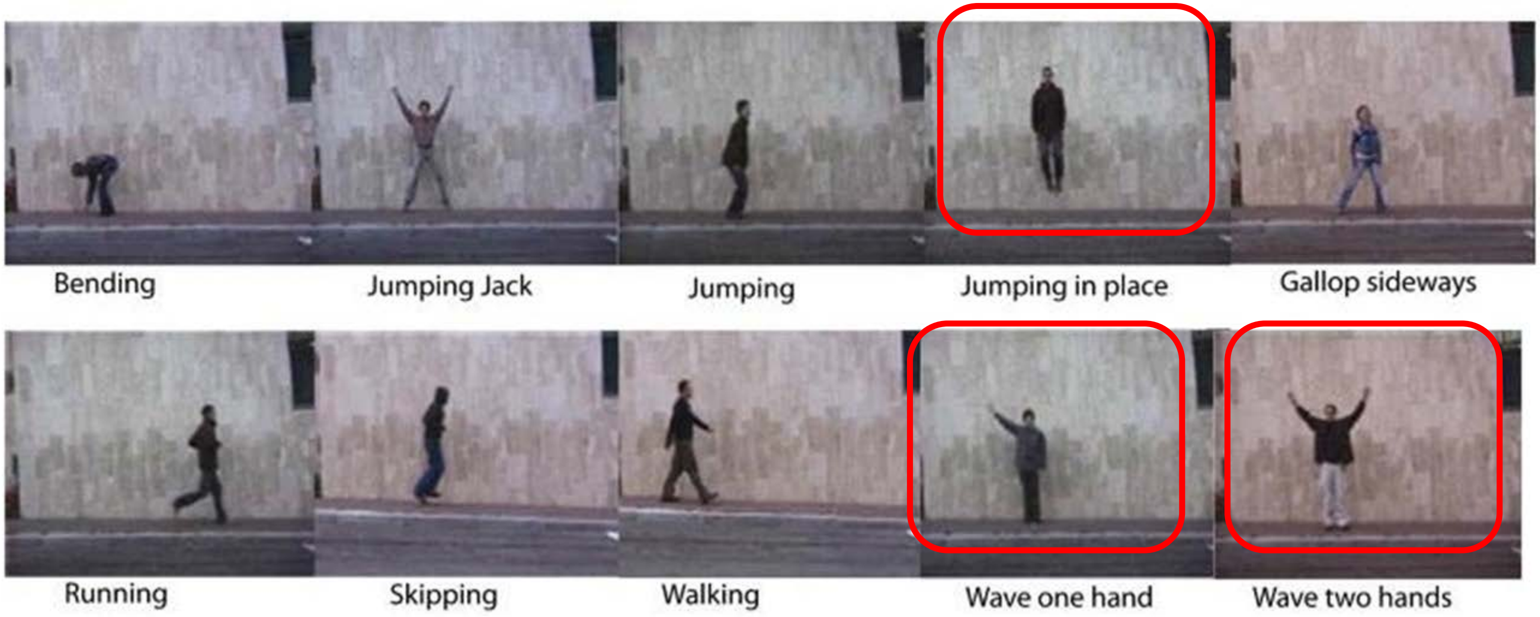
Accuracy in recognizing behavior categories



(Jung, Hwan & Tani, IEEE ICDL-EPIROB 2014)

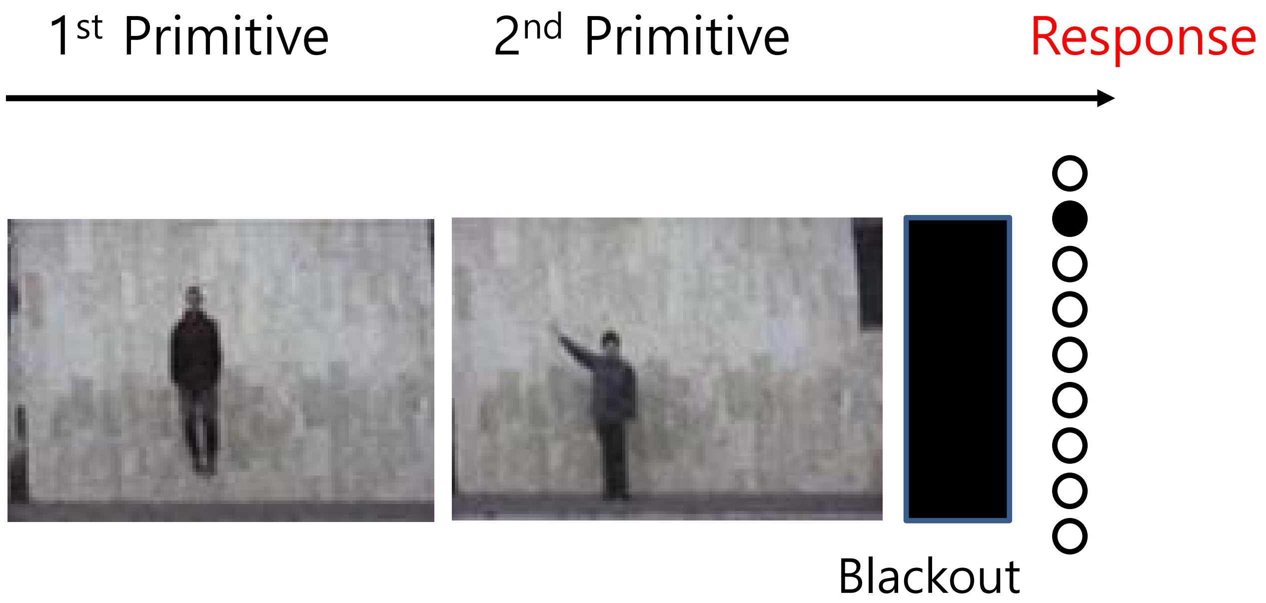
Categorization of Compositional Action Sequences

(Jung, Hwan & Tani, IEEE ICDL-EPIROB 2014)



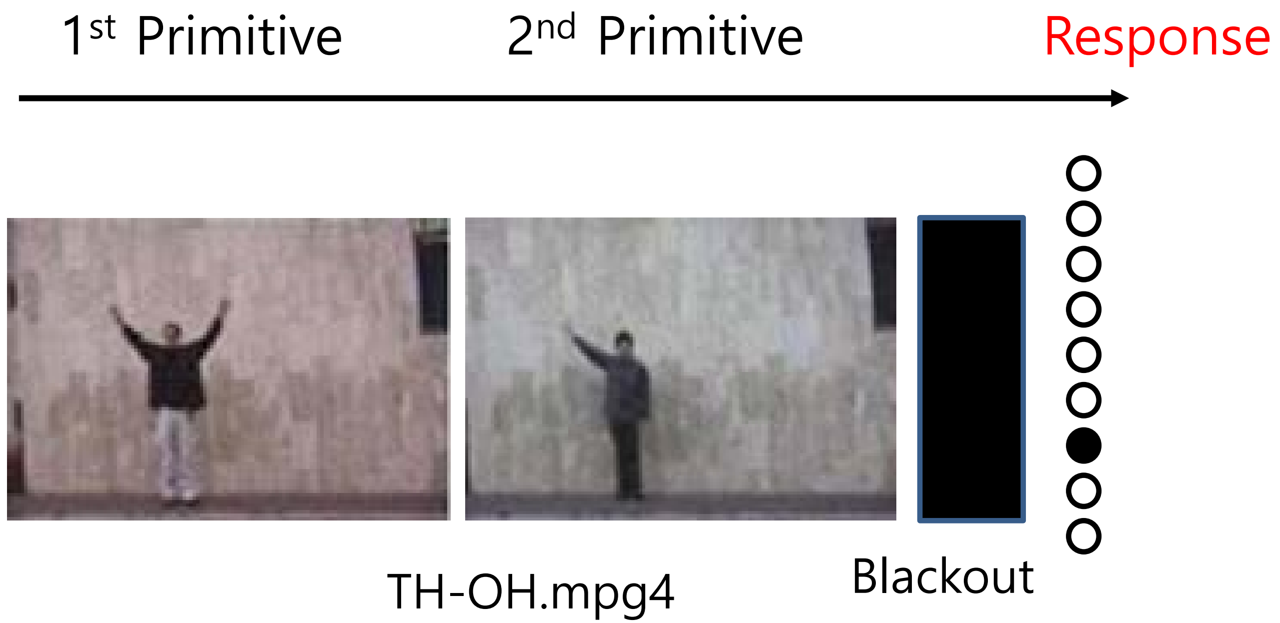
- Use 3 primitive patterns from Weizmann dataset.
- Then, construct sequential combination of them for learning.

Categorization of Action Sequences



JP-OH.mpg4

Categorization of Action Sequences



- Leave one subject cross validation with 9 subjects
  - Training with 8 out of 9 subjects data, testing with the untrained subject data and repeat it 9 times for averaging.

VIDEO-2

Effects of Slow Timescale

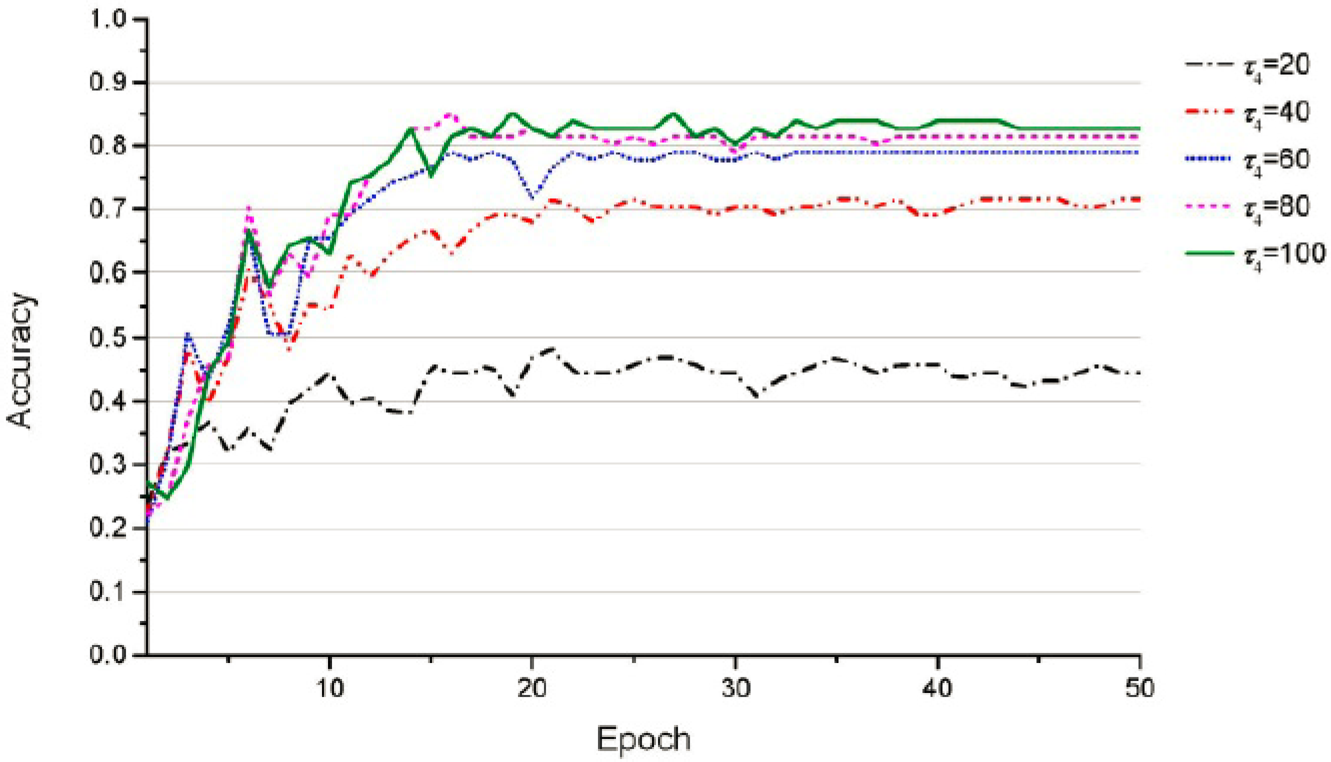
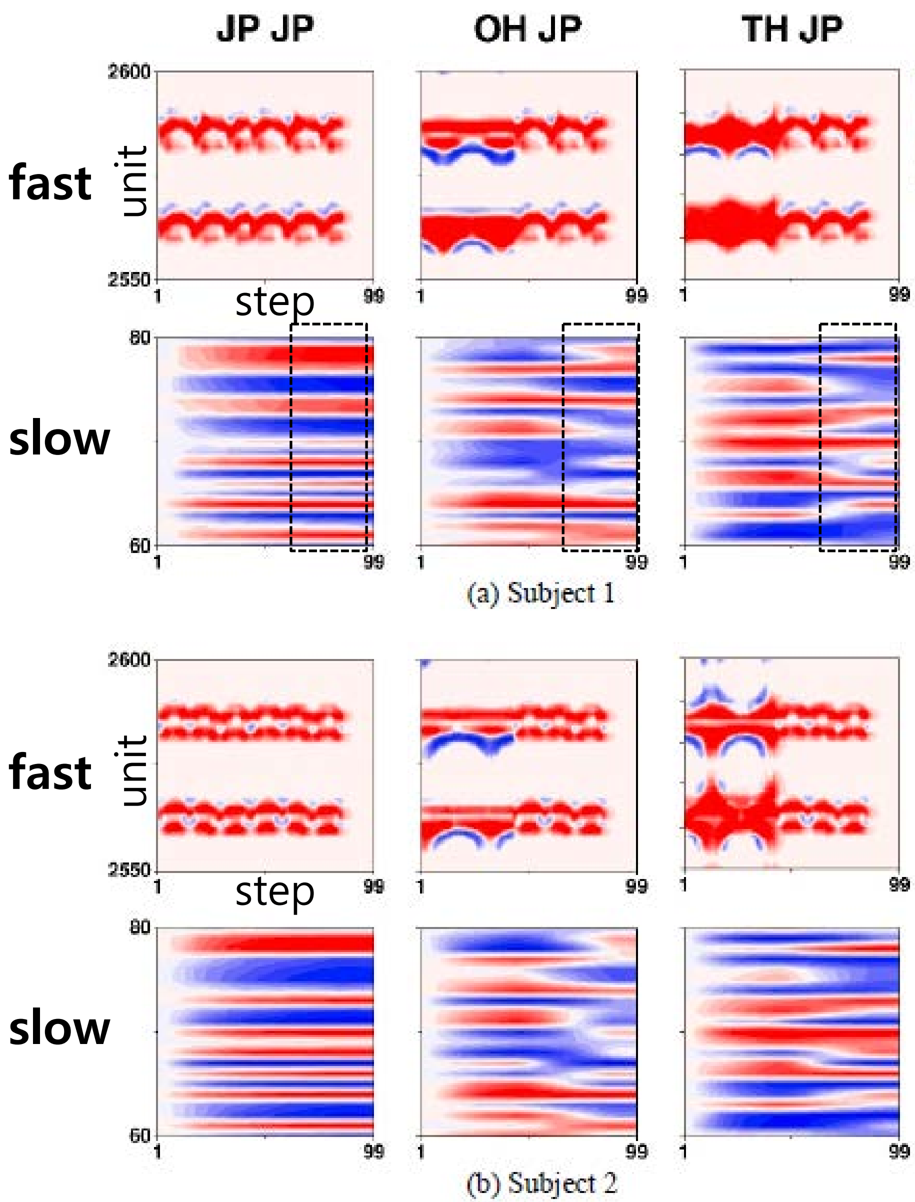


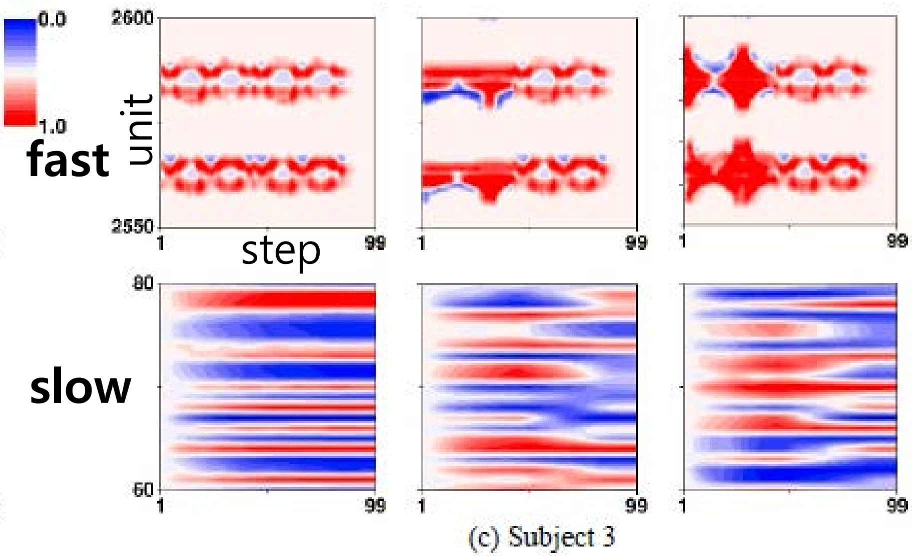
Fig. 5. Development of recognition accuracy with different time constants assigned for the layer 4. The vertical axis represents the accuracy obtained from leave-one-subject-out cross-validation and horizontal axis represents epochs during training phase. By changing the time constant from small ($\tau_4 = 20.0$) to large ($\tau_4 = 100.0$) stepping by 20, the accuracy is largely increased.

Contextual Process by Each Subject



(a) Subject 1

(b) Subject 2



(c) Subject 3

(Jung, Hwan & Tani, 2015)

Achieving “Synergy” in Visuo-Motor-Attentional Coordination

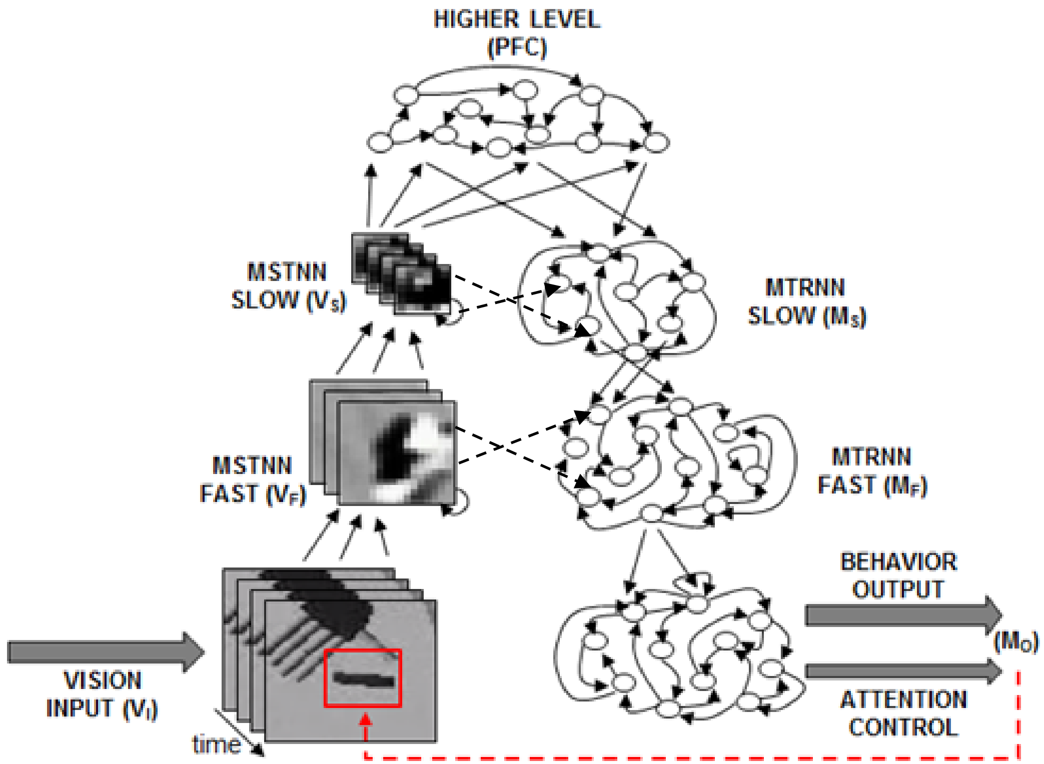
Jungsik Hwang1, Minju Jung1, Naveen Madapana2, Jinhyung Kim1, Minkyu Choi1 and Jun Tani1* (IEEE Humanoid Robots 2015)

“Synergetic” Coordination among Multiple Cognitive Processes in Human Interaction Tasks

- A higher-order cognitive action requires adequate coordination among multiple cognitive processes.
  - Visual recognition (human gestures, own movements, objects)
  - Attention shifts
  - Action sequence preparation
  - Visuo-motor coordination

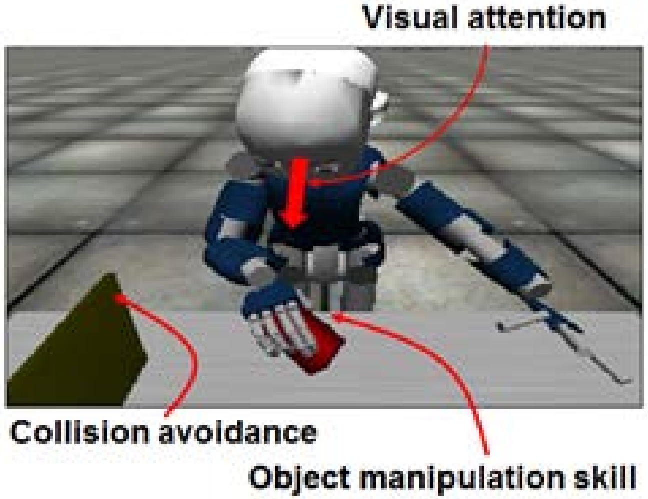
Visuo-Motor Deep Dynamic Neural Network (VMDNN)

(Jungsik Hwang et al, 2015)

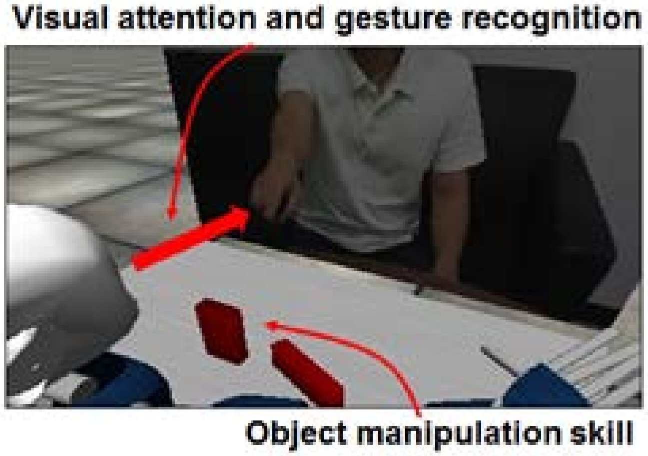


Cognitive Behavioral Tasks

Taks-1



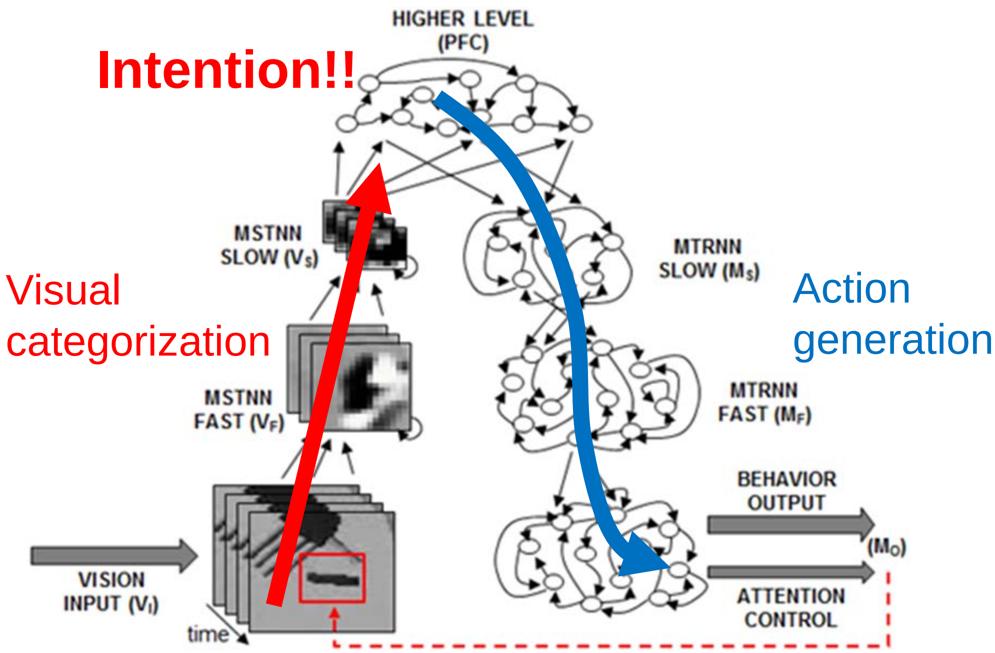
Taks-2

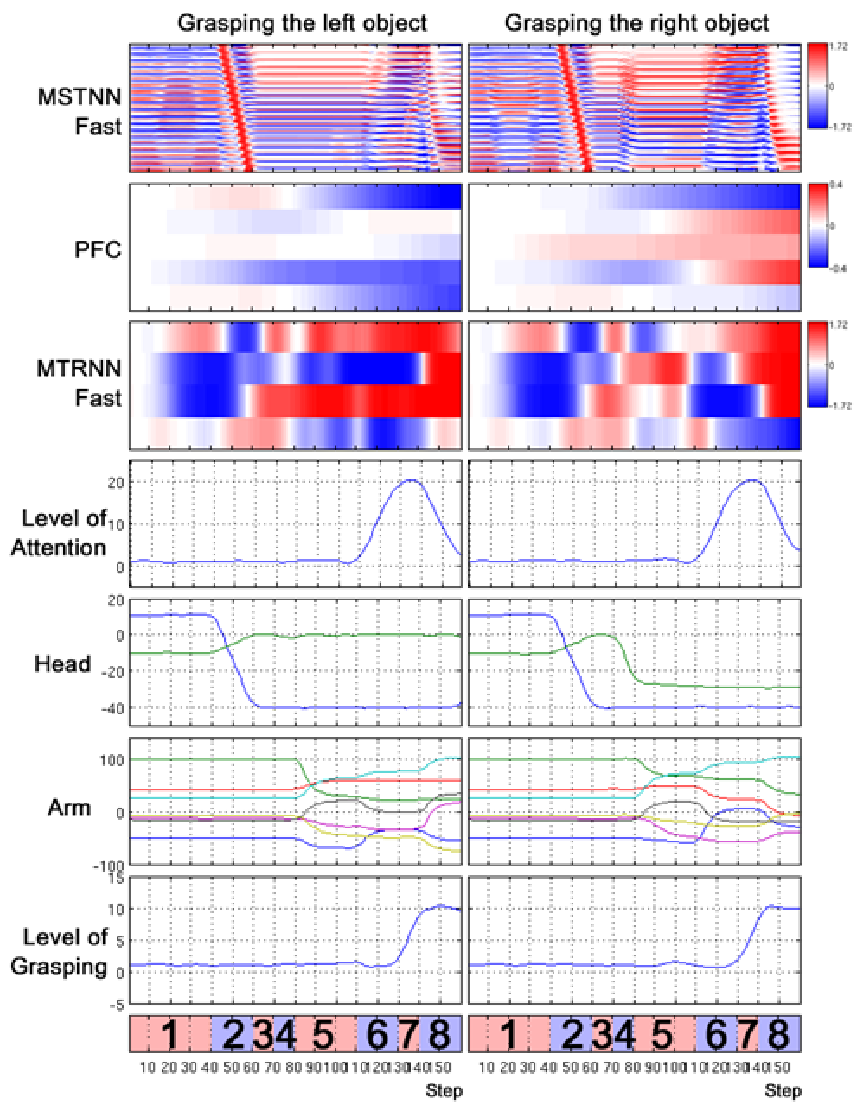


Video

Video

Neural Dynamics in Task-2





1. Observing human gestures.
2. Attending to the task space.
3. Observing the task space.
4. Attending to the target object.
5. Reaching the hand to the above of the object.
6. Reaching the hand to the near the object with focusing it.
7. Grasping it.
8. Lifting it.

Summary

- Future humanoid robots would depend more on learning instead of programming.
- Functional hierarchy in generating action and in recognizing dynamic visual pattern can be developed via learning of visuo-motor pattern when spatio-temporal constraints are adequately imposed on the network dynamics.
- Synergy among different cognitive processes can be achieved by allowing dense interactions among subnetworks.
- Future studies should challenge more complex robotic tasks in social cognitive contexts.

Collaborators in Cog. Neuro-Robotics Lab in KAIST

Ph.D students
Minju Jung
Haanvid Lee
Minkyu Chi
Jungsik Hwan
Gibeom Park

This research has been supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (No. 2014R1A2A2A01005491).