

Deep Affordance-Grounded Sensorimotor Object Recognition

Authors: Spyridon Thermos, Georgios
Th. Papadopoulos, Petros Daras,
Gerasimos Potamianos

Presented By:
Thomas Crosley
UT CS 381V
Autumn 2017



Problem

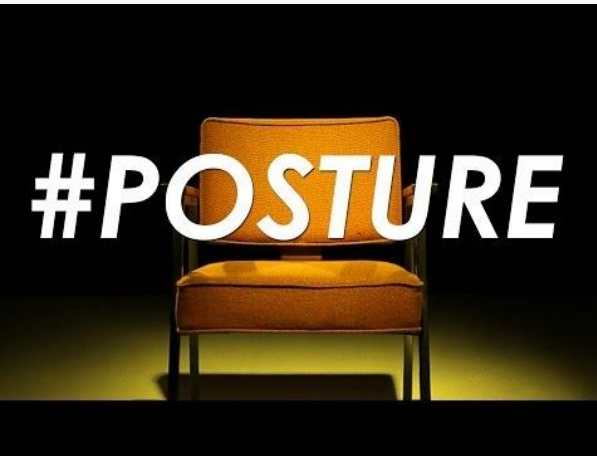
- Integrate visual **appearance** and visual **affordance** information
- Object + Affordance Classification



Hit Using Hammer

Affordances: “the types of actions that humans typically perform when interacting with an object.”

Sit



<https://www.youtube.com/watch?v=1xS864zYlo8>

Throw



<https://www.youtube.com/watch?v=7Qxu5cvW-ds>

Workout



<https://www.youtube.com/watch?v=V4XW74W9t4o>

Smaller Data

- Few objects [1, 2, 3]
- Small number of affordances [1, 2, 3]
- Ex: 6 objects and 3 affordances [1]

Simpler Methods

- Factorial Conditional
Random Fields and Binary
SVMs [1]
- Gaussian Processes [2]
- SVMs + Clustering [3]

[1]

[2]

[3]

RGB-D Sensorimotor Dataset





CERTH
CENTRE FOR
RESEARCH & TECHNOLOGY
HELLAS



UNIVERSITY OF
THESSALY

RGB-D Sensorimotor Dataset

Step 1

RGB (1920x1080)



Mapping

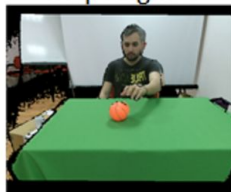


Mapping

Depth Map (512x424)

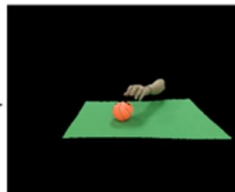
Step 2

RGB & Depth
Map Aligned



512x424

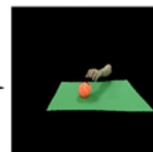
3D Volume of
Interest Cropped



512x424

Step 3

Cropped



300x300

Hand & Object
Segmentation

Hand & Object
Segmentation

RGB (300x300)



Hand



Object



Depth Maps (300x300)



RGB-D Sensorimotor Dataset

Step 1

RGB (1920x1080)



Mapping

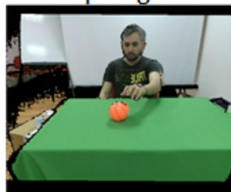


Mapping

Depth Map (512x424)

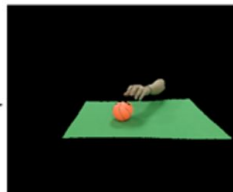
Step 2

RGB & Depth
Map Aligned



512x424

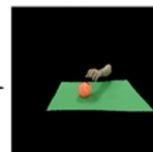
3D Volume of
Interest Cropped



512x424

Step 3

Cropped



300x300

Hand & Object
Segmentation

Hand & Object
Segmentation

RGB (300x300)



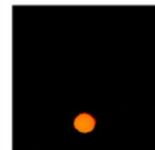
Hand



Object



Depth Maps (300x300)



Original Input

RGB-D Sensorimotor Dataset

Step 1

RGB (1920x1080)



Mapping

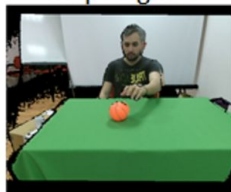


Mapping

Depth Map (512x424)

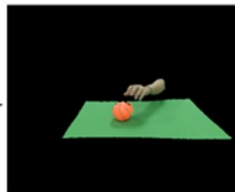
Step 2

RGB & Depth
Map Aligned



512x424

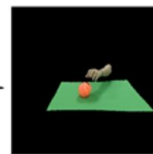
3D Volume of
Interest Cropped



512x424

Step 3

Cropped



300x300

Hand & Object
Segmentation

Hand & Object
Segmentation

RGB (300x300)



Hand



Object



Depth Maps (300x300)



Input Processing

RGB-D Sensorimotor Dataset

Step 1

RGB (1920x1080)



Mapping

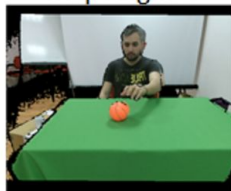


Depth Map (512x424)

Mapping

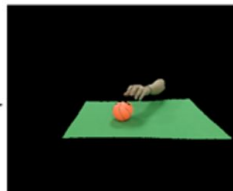
Step 2

RGB & Depth
Map Aligned



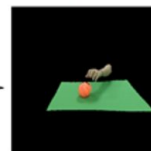
512x424

3D Volume of
Interest Cropped



512x424

Cropped



300x300

Step 3

Hand & Object
Segmentation

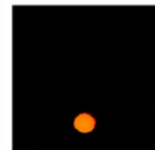
Hand & Object
Segmentation

RGB (300x300)



Hand

Object

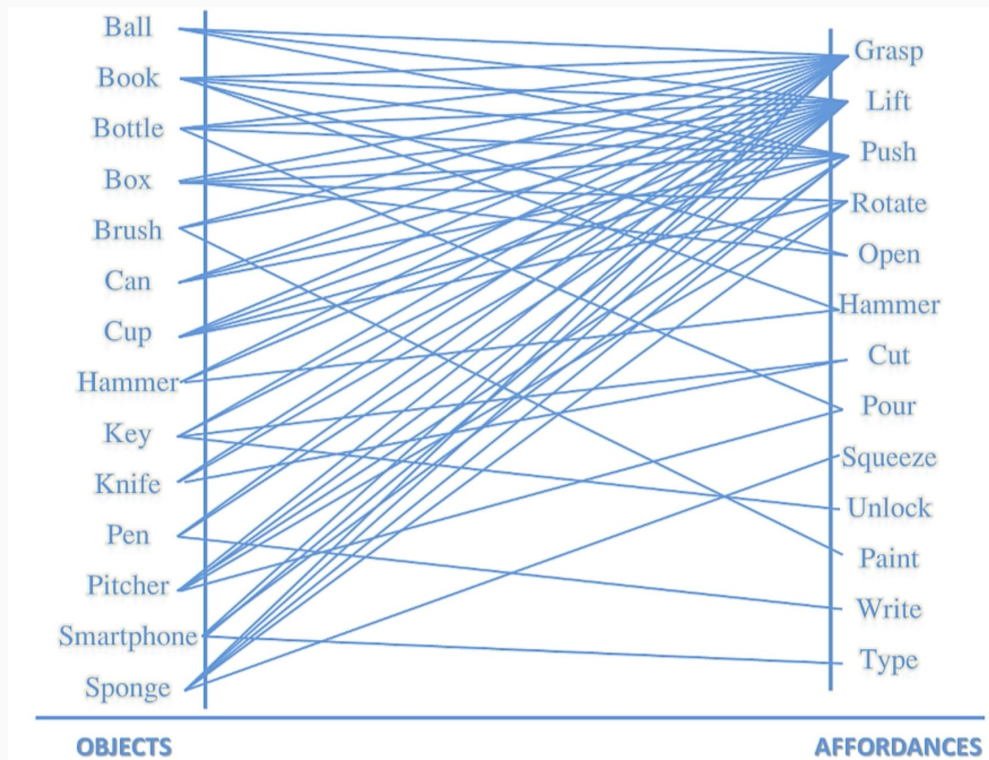


Depth Maps (300x300)

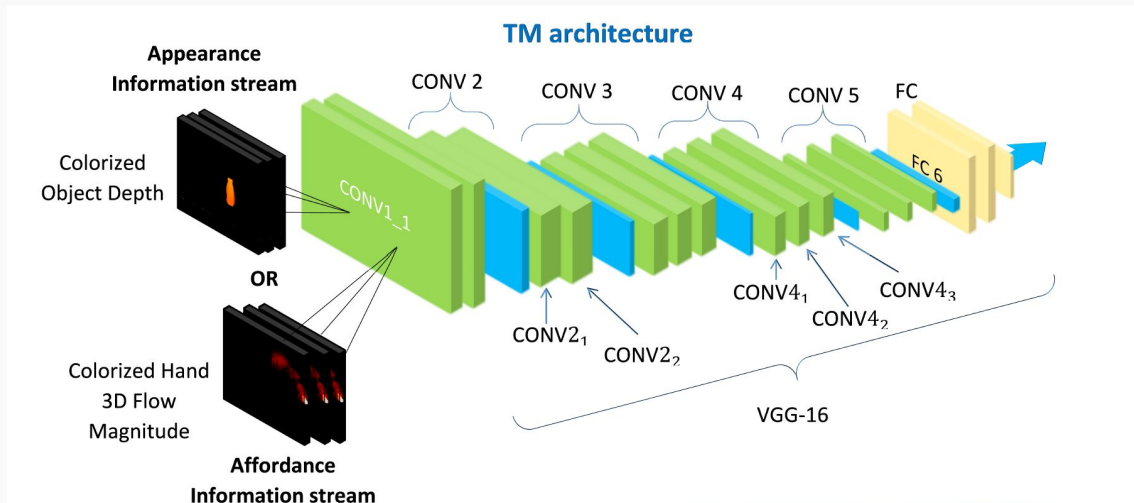
Data Extraction

RGB-D Sensorimotor Dataset

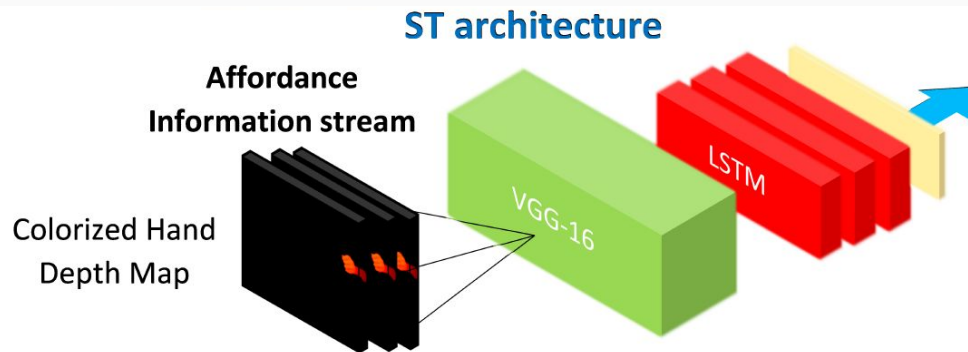
- 14 Object Types
- 13 Affordances
- 54 Interactions
- 105 subjects
- 4 to 8 seconds
- 20,830 instances



Architectures



- Generalized Template-Matching (GTM)
- Model spatial correlations
- Appearance CNN for object detection



- Generalized Spatio-Temporal (GST)
- Encode time-evolving procedures
- CNN+LSTM for affordance modeling

Long Short Term Memory Networks (LSTMs)

LSTMs: recurrent architecture capable of learning long-term dependencies

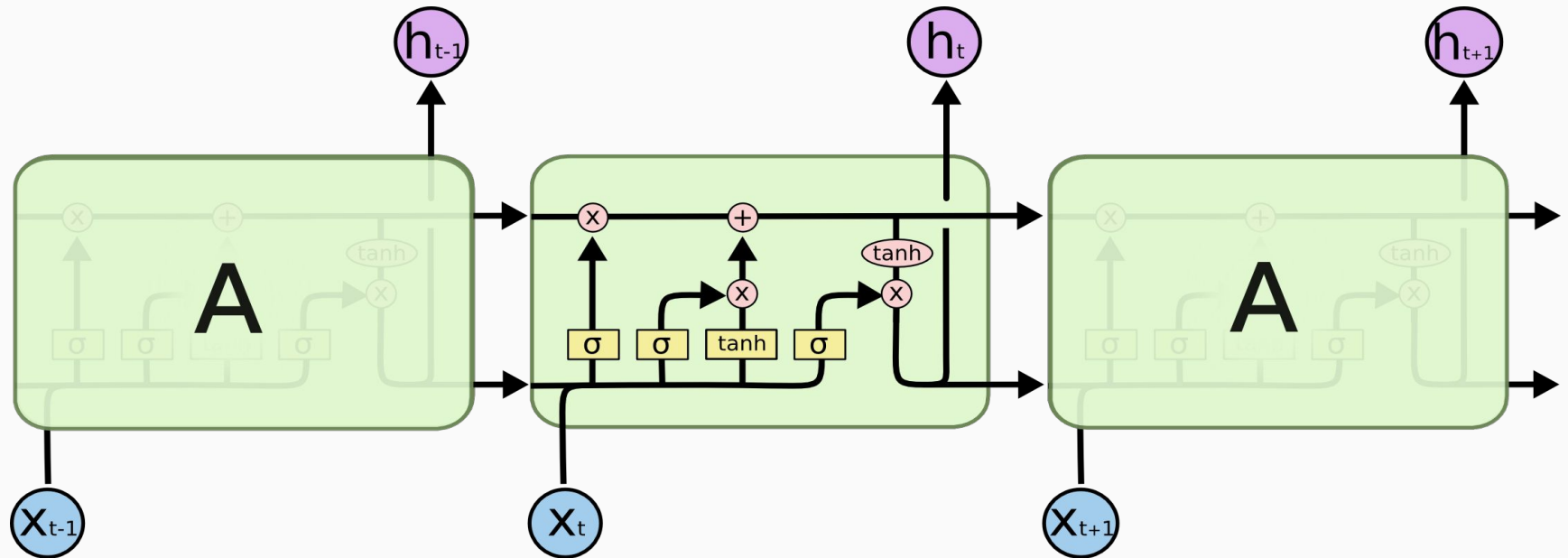
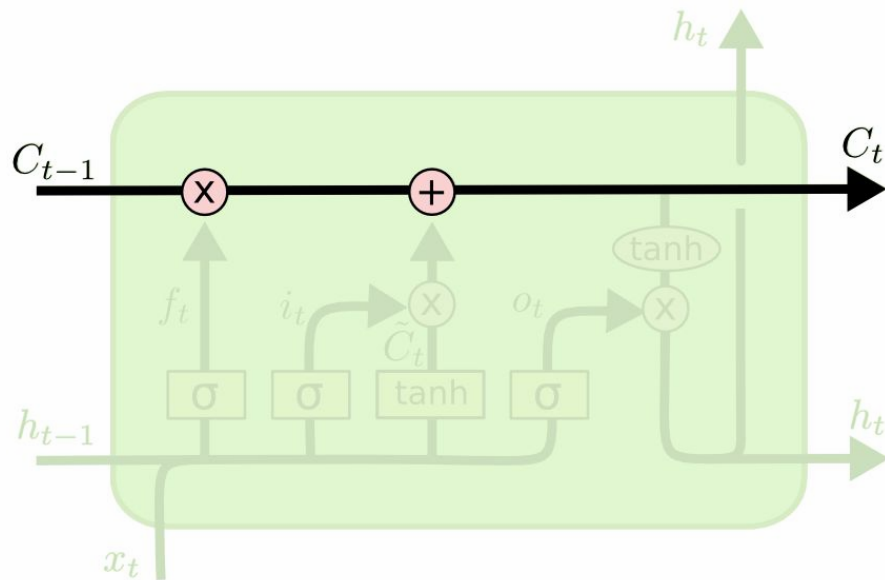
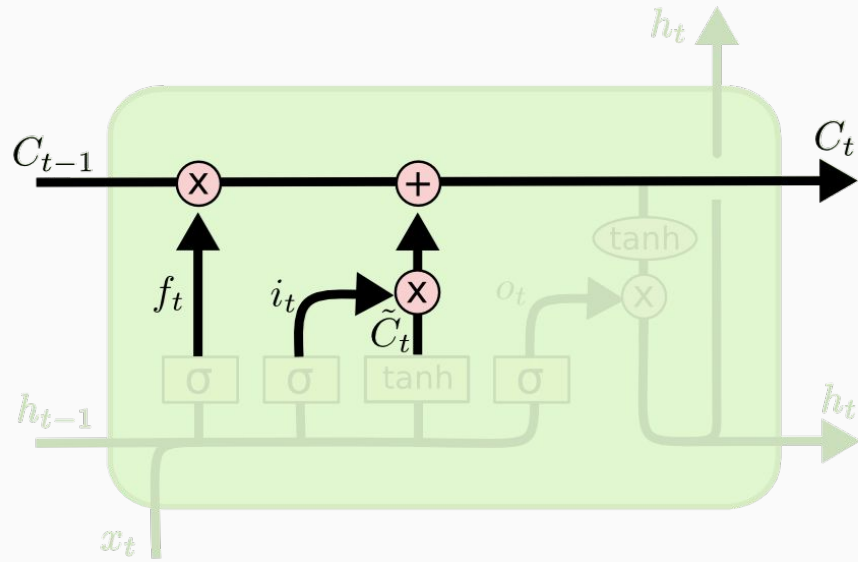


Image Source: <http://colah.github.io/posts/2015-08-Understanding-LSTMs/>

Core Idea: cell state updated and then passed on at each time step



LSTMs

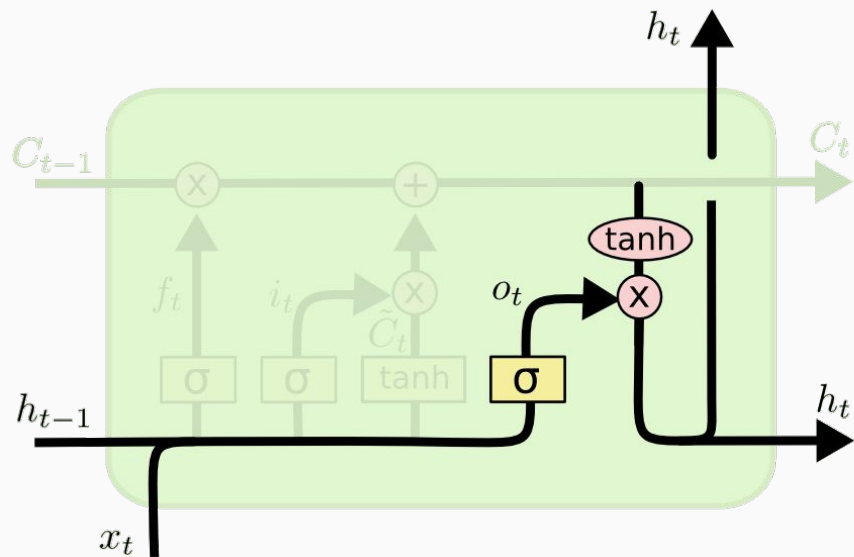


$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

“Forget Gate”

“Remember Gate”

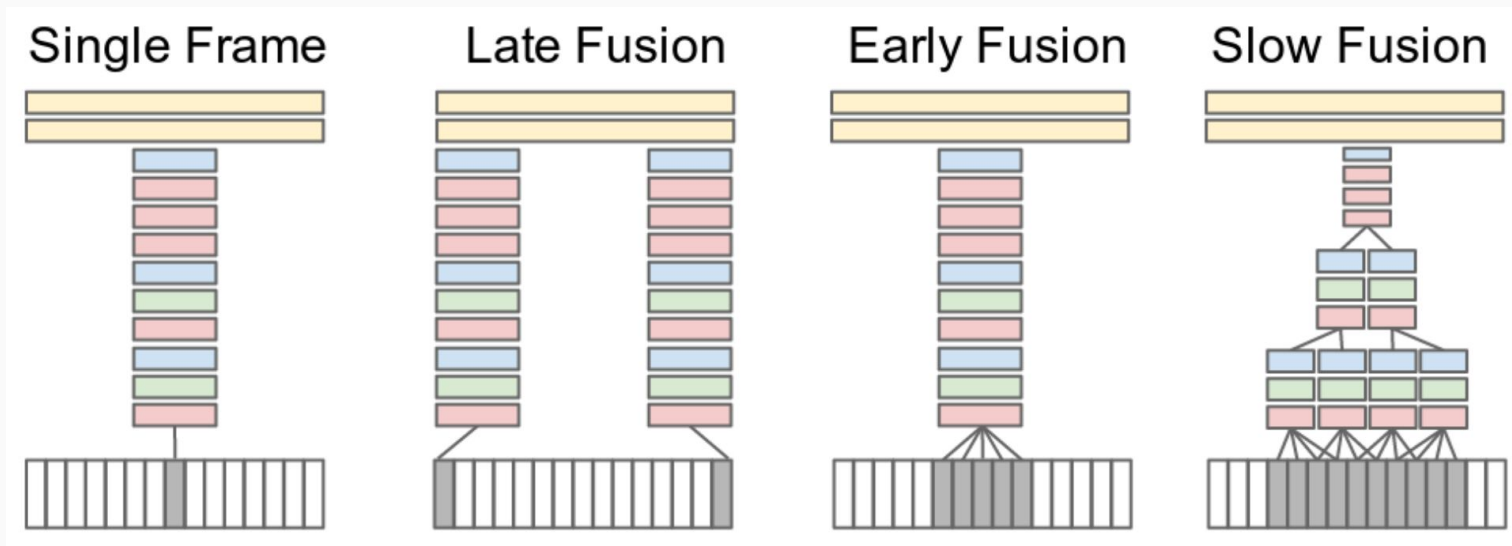
LSTMs



$$o_t = \sigma (W_o [h_{t-1}, x_t] + b_o)$$

$$h_t = o_t * \tanh (C_t)$$

Fusion

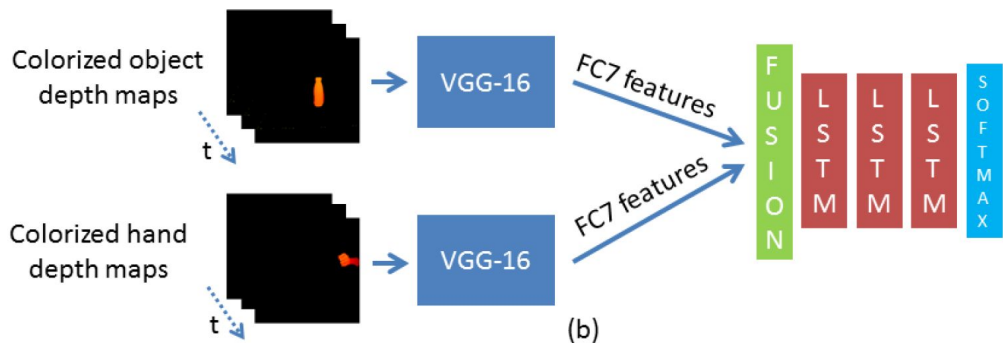
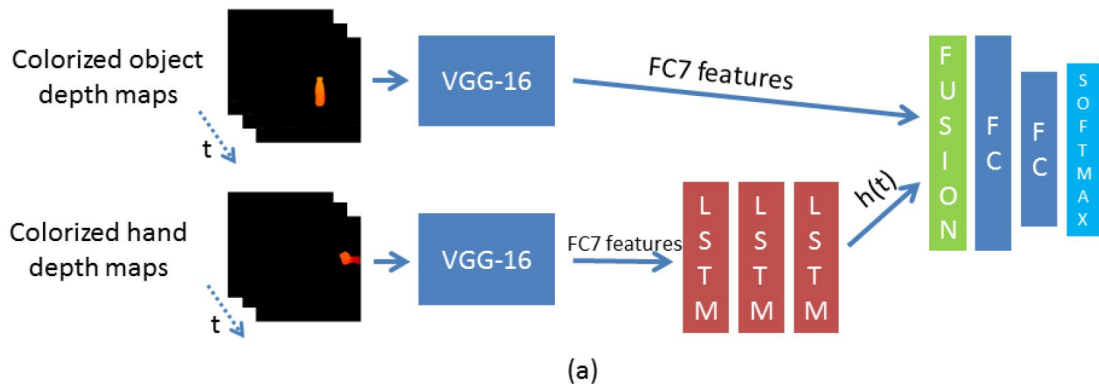


- Given multiple sources of information
- At what point do we combine their features?

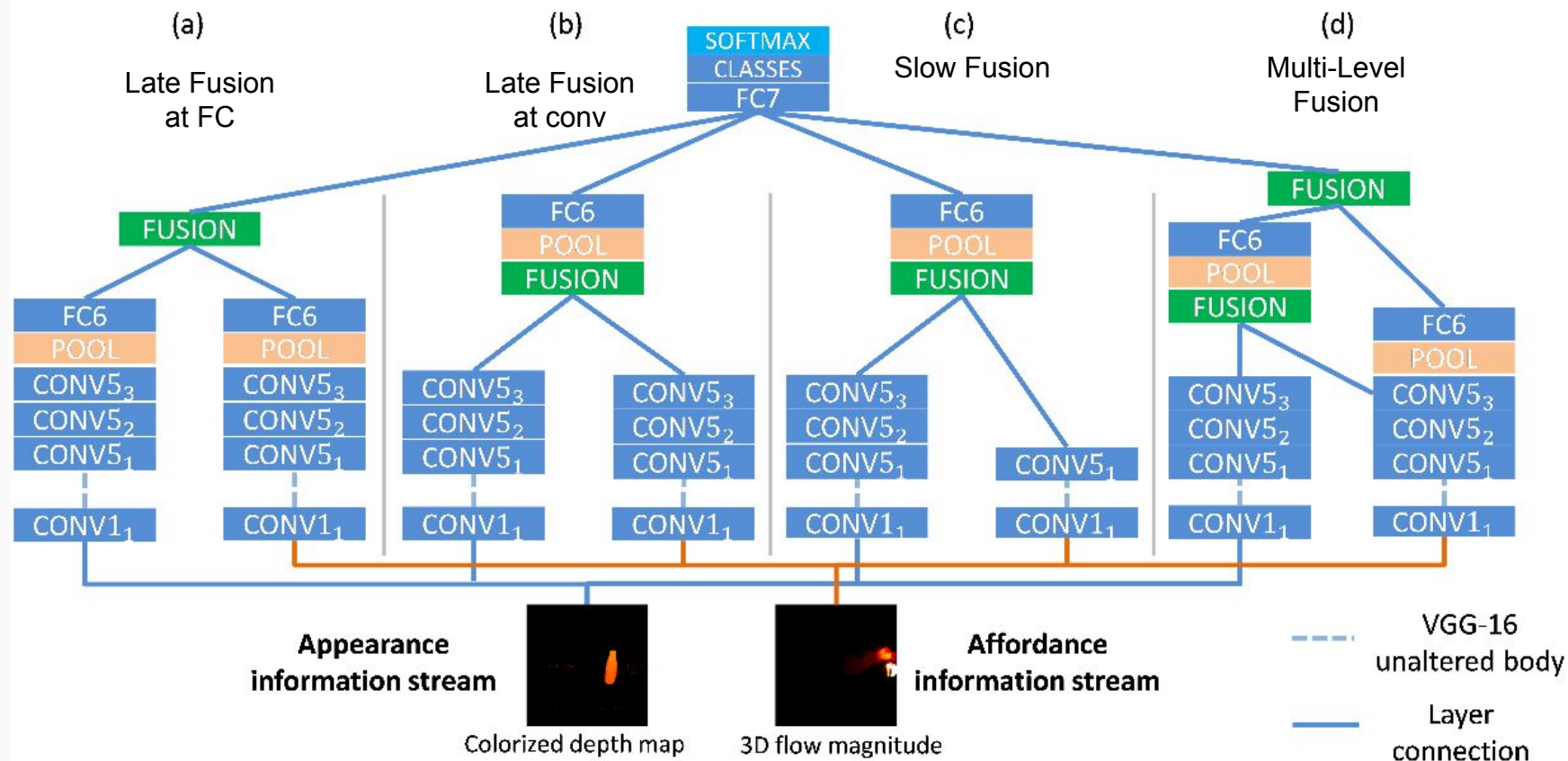
Image Source: <http://cs.stanford.edu/people/karpathy/deepvideo/>

Fusion

- GST Architecture
- Combines
 - Appearance
 - Affordance
- (a) Late Fusion
- (b) Slow fusion



Architecture



Results

Single Stream

Method	Task	Accuracy (%)
Appearance CNN	object recognition	85.12
Affordance CNN	affordance recognition	81.92
Affordance CNN-LSTM	affordance recognition	69.27

(Best) Template Matching

GTM-based fusion architecture [after fusion]	Accuracy (%)
$GTM_{SML}(RL5_3^{app}, RL5_1^{aff}, RL6)$	88.23
$GTM_{SML}(RL5_3^{app}, RL5_3^{aff}, RL6)$	89.43

(Best) Spatio-Temporal

GST-based fusion architecture [after fusion]	Accuracy (%)
$GST_{LS}(\text{last-frame})$	86.28
$GST_{LS}(\text{all-frames})$ [1 CONV, 2 FC]	86.50

- Authors' Thoughts
 - NN-Autoencoders for human-object interactions
 - “In-the-wild” object-affordance detection
- Others
 - Affordance identification for control tasks
 - Better temporal sampling schemes