UNIVERSITAT DE BARCELONA

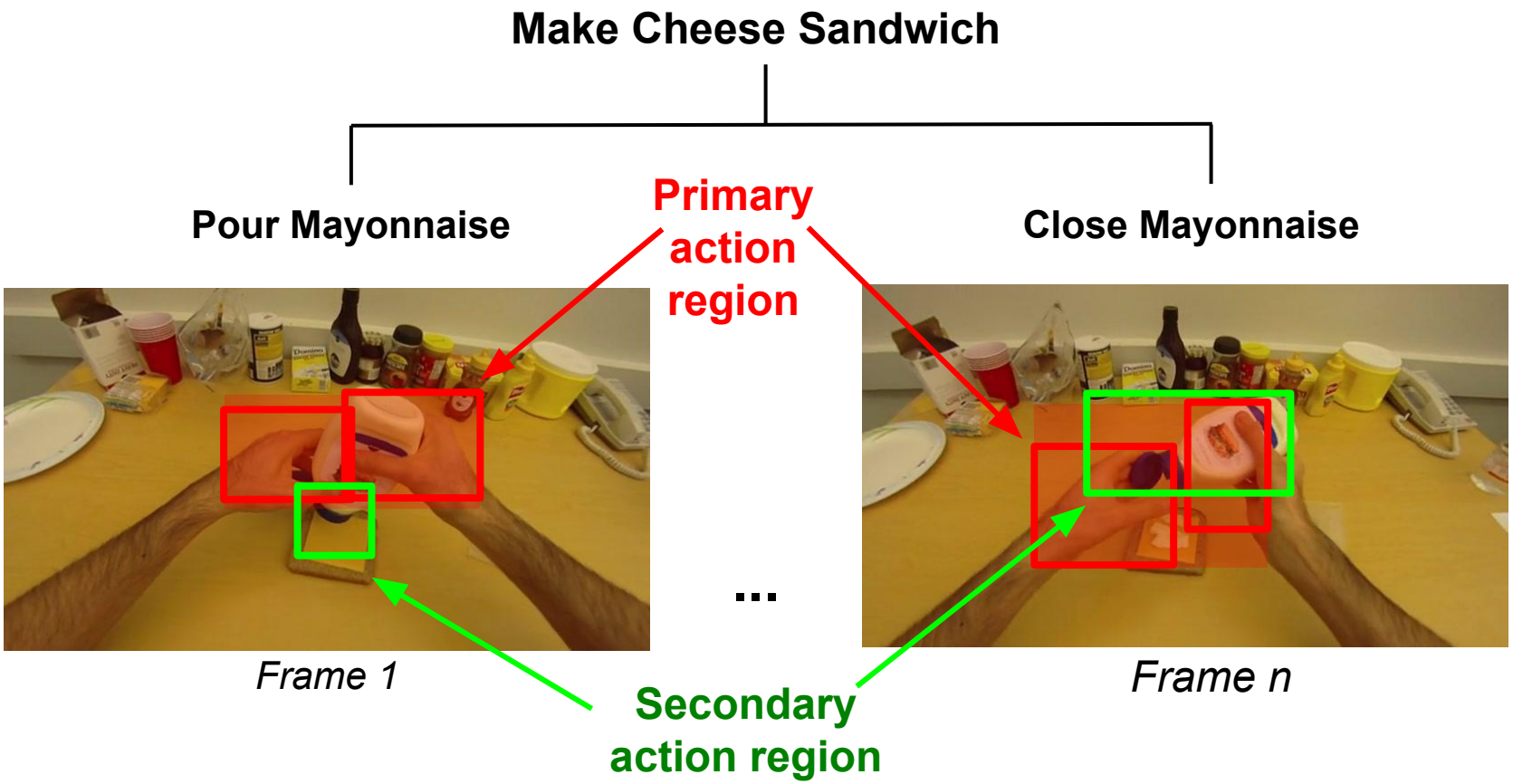

# Modeling Long-Term Interactions to Enhance Action Recognition

January 2021

**Alejandro Cartas Ayala**, Petia Radeva, and Mariella Dimiccoli

# Motivation | Spatio-Temporal Action Regions

- We model **context regions** and their **temporal evolution**.



# Previous work

- First deep architecture was proposed by **Ma et al. [2016]**.
- Contextual action recognition was proposed by **Gkioxari et al. [2015]**.

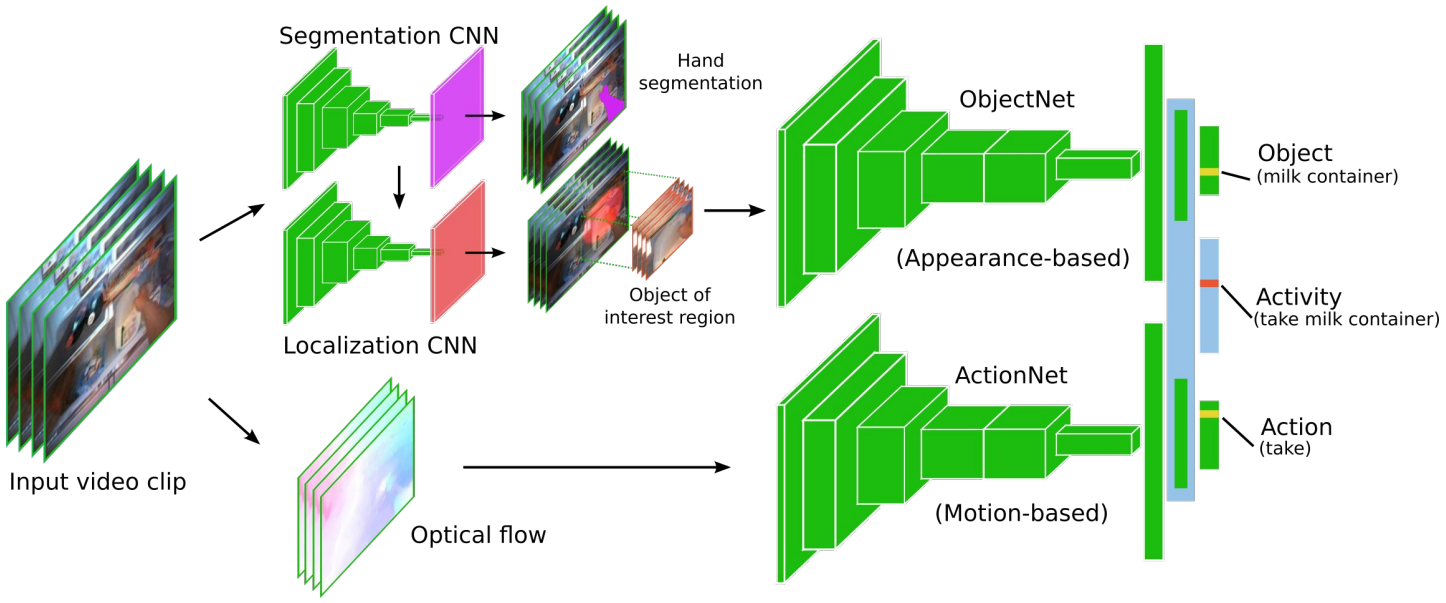


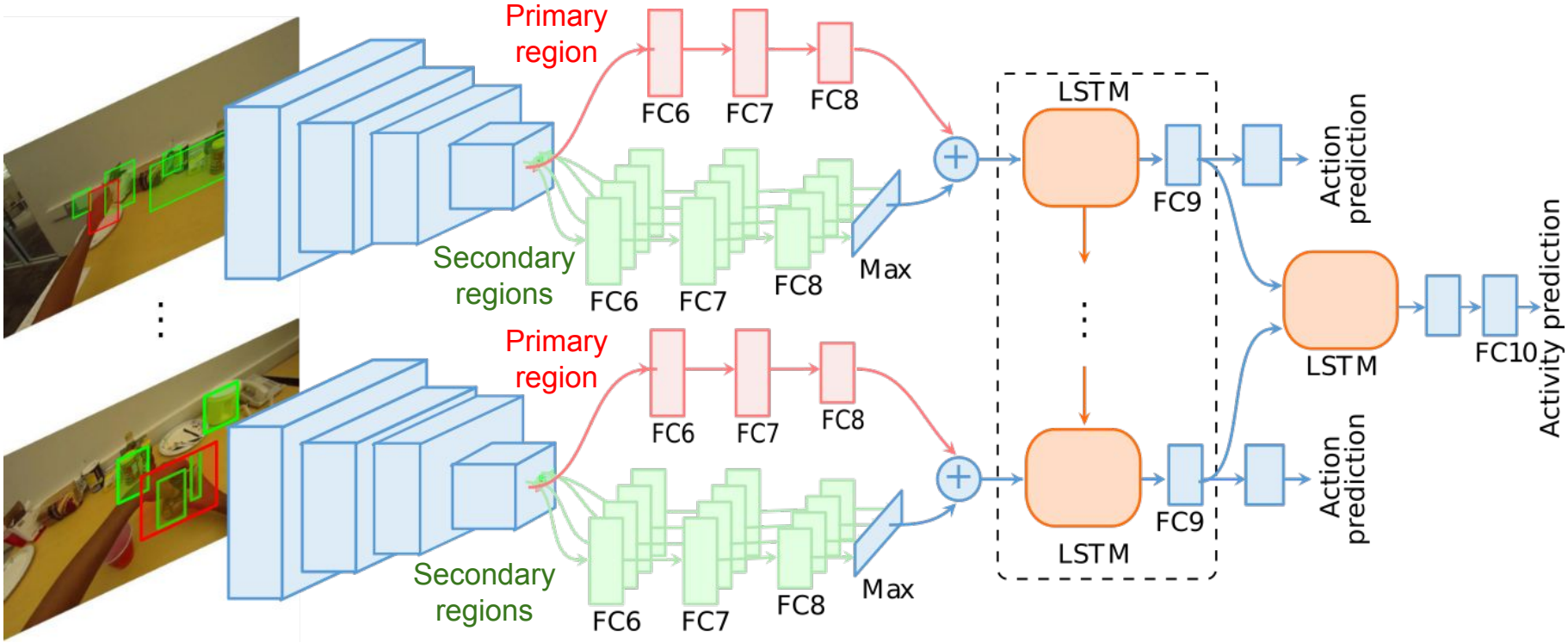

Image taken from *Ma et al.* [2016]

# Goals

- To present a region-based architecture that:
  - Models contextual relationships.
  - Predicts **Action** and **Activities**.

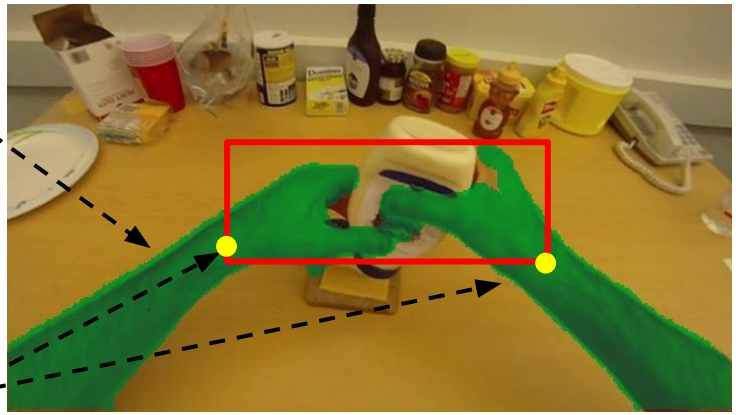
- To present **two data augmentation strategies**:
  - Visual augmentation.
  - Sequence augmentation.

# Methodology | Region-Based Action Recognition Method



# Methodology | Primary Action Region: Hands

**1. Skin detection**
(Li and Kitani 2013).

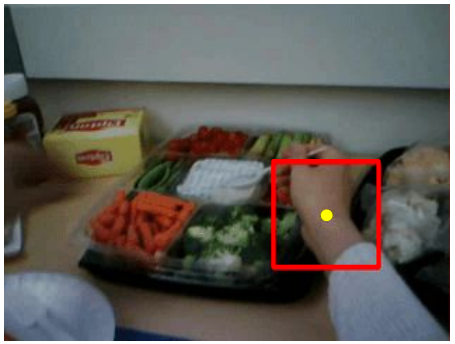


**2. Wrist location**
PAFs (Cao et al. 2017).

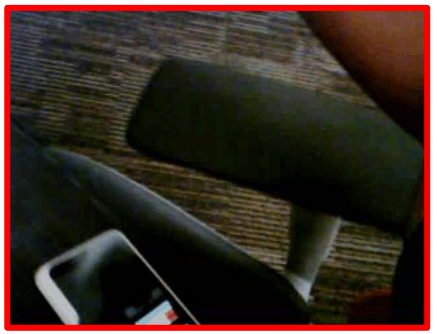
**3. Rules for special cases:**

No wrist is located

No skin is detected



No skin or wrist is found

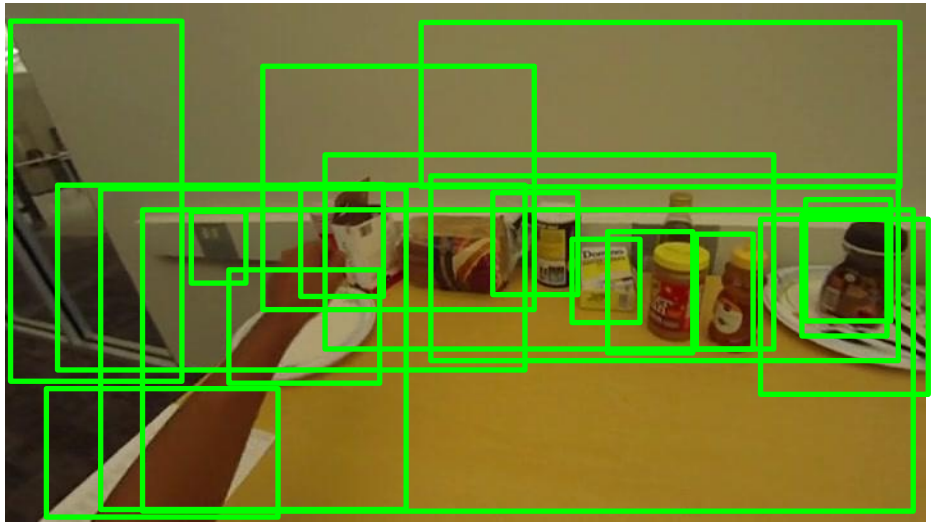


# Methodology | Secondary Action Region

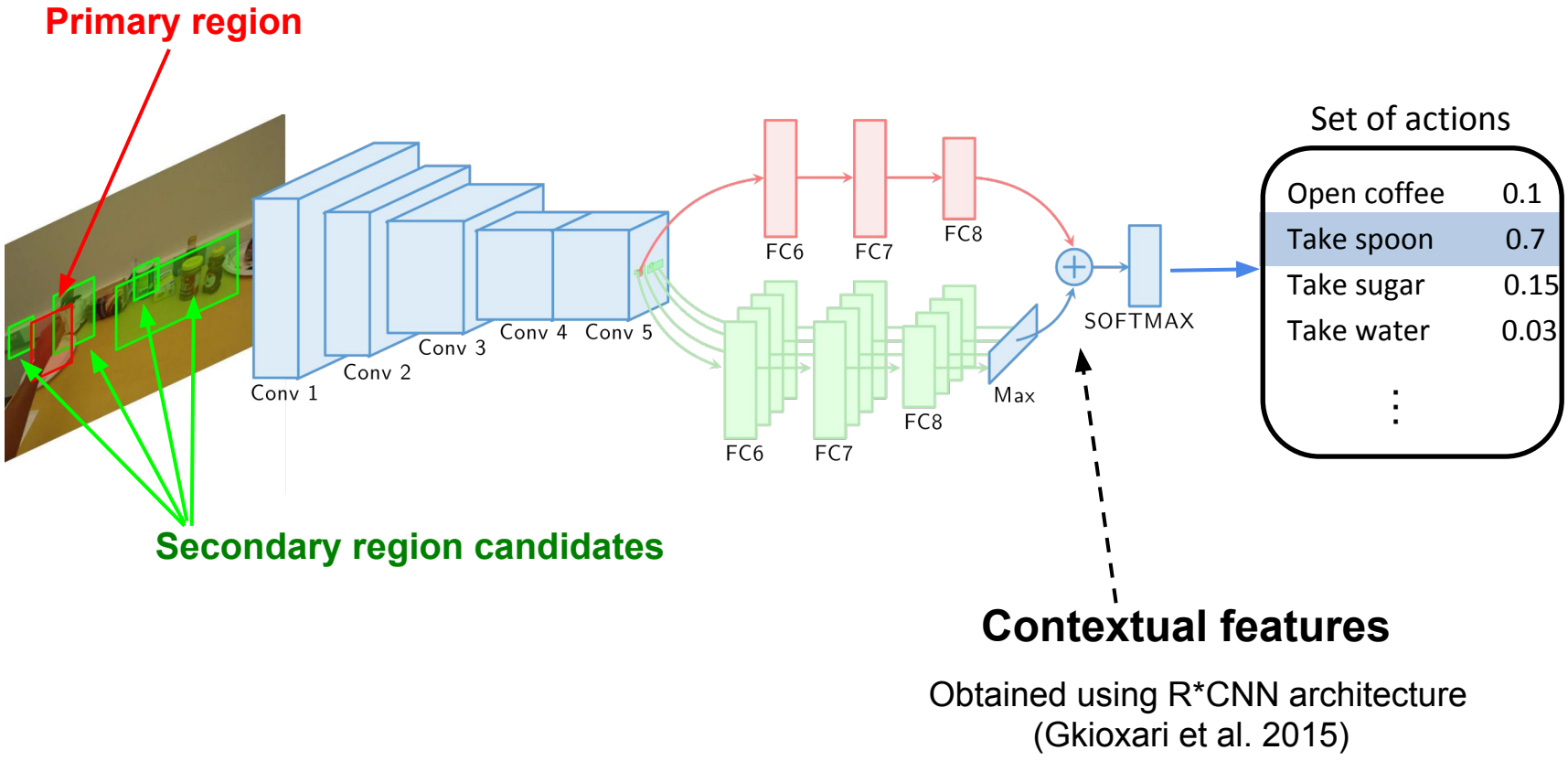
- A region is chosen after classification.
- Up to 2K regions are proposed as candidates.
  - Selective Search (**Uijlings et al., 2013**)
  - MCG (**Arbelaez et al. 2014**)

# Methodology | Secondary Action Region

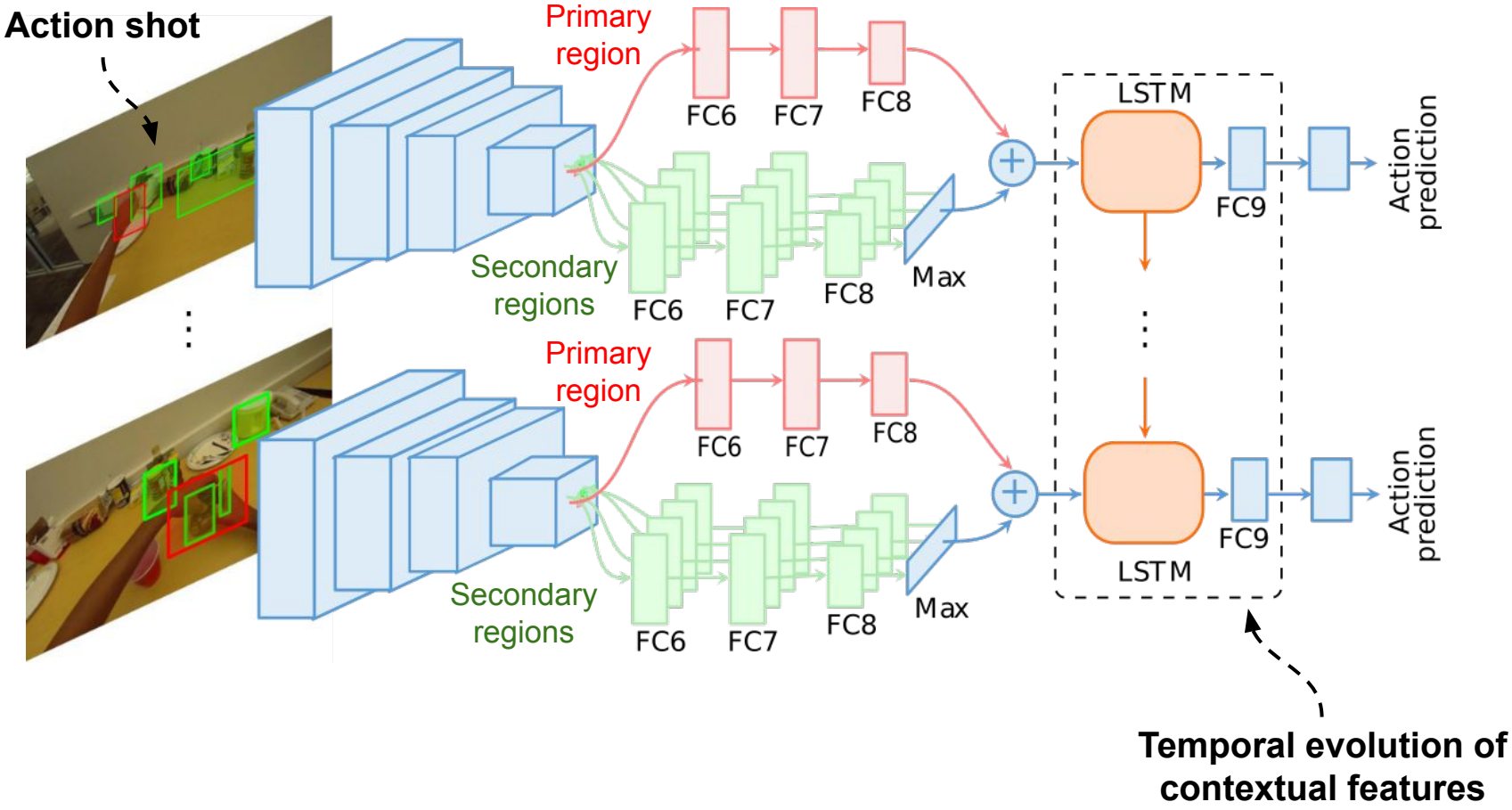
- Up to 2K regions are proposed as candidates.
  - Selective Search (**Uijlings et al., 2013**)
  - MCG (**Arbelaez et al. 2014**)
- A region is chosen after classification.



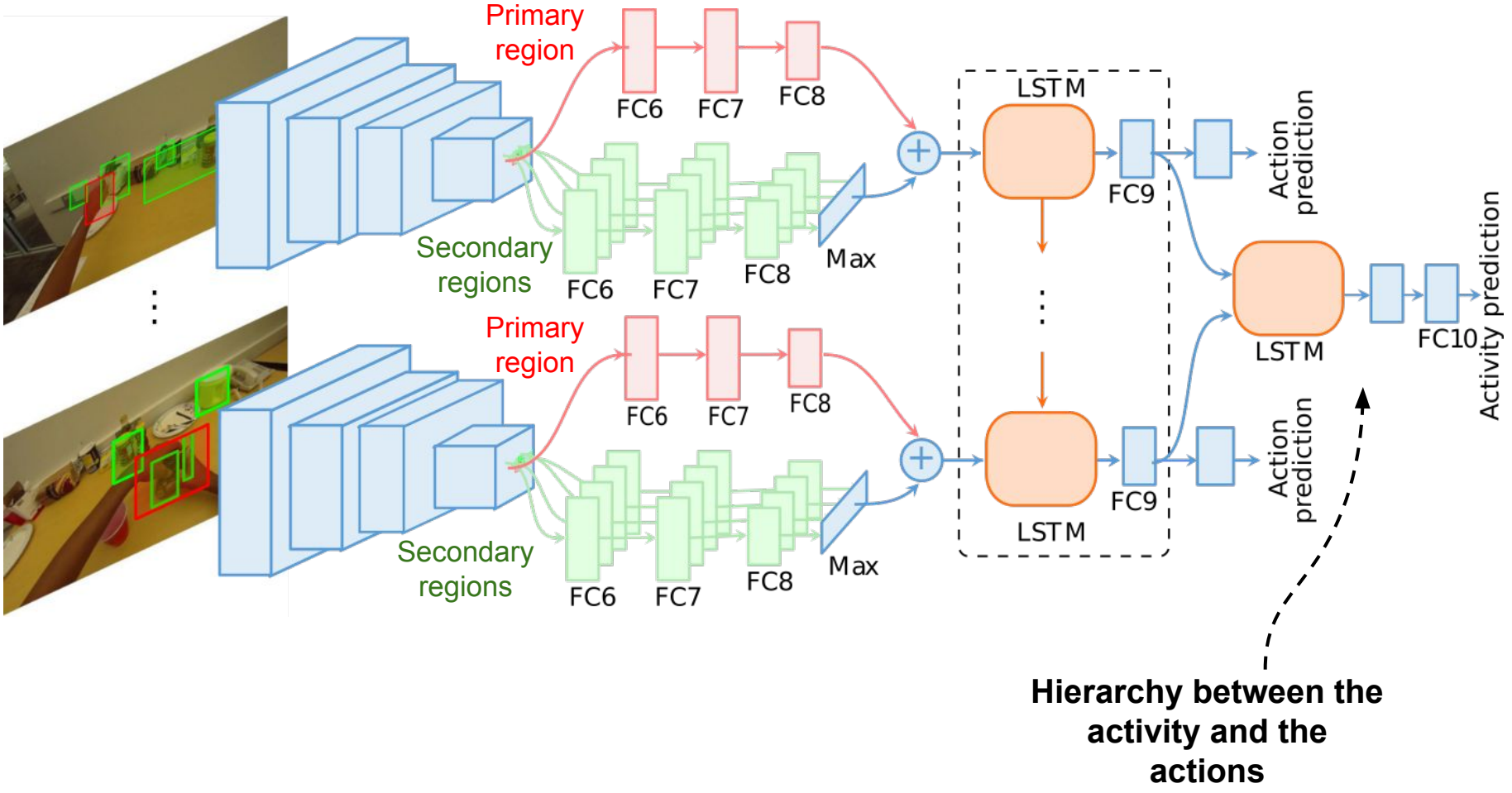
# Methodology | Frame-Level Modeling



# Methodology | Action-level Modeling

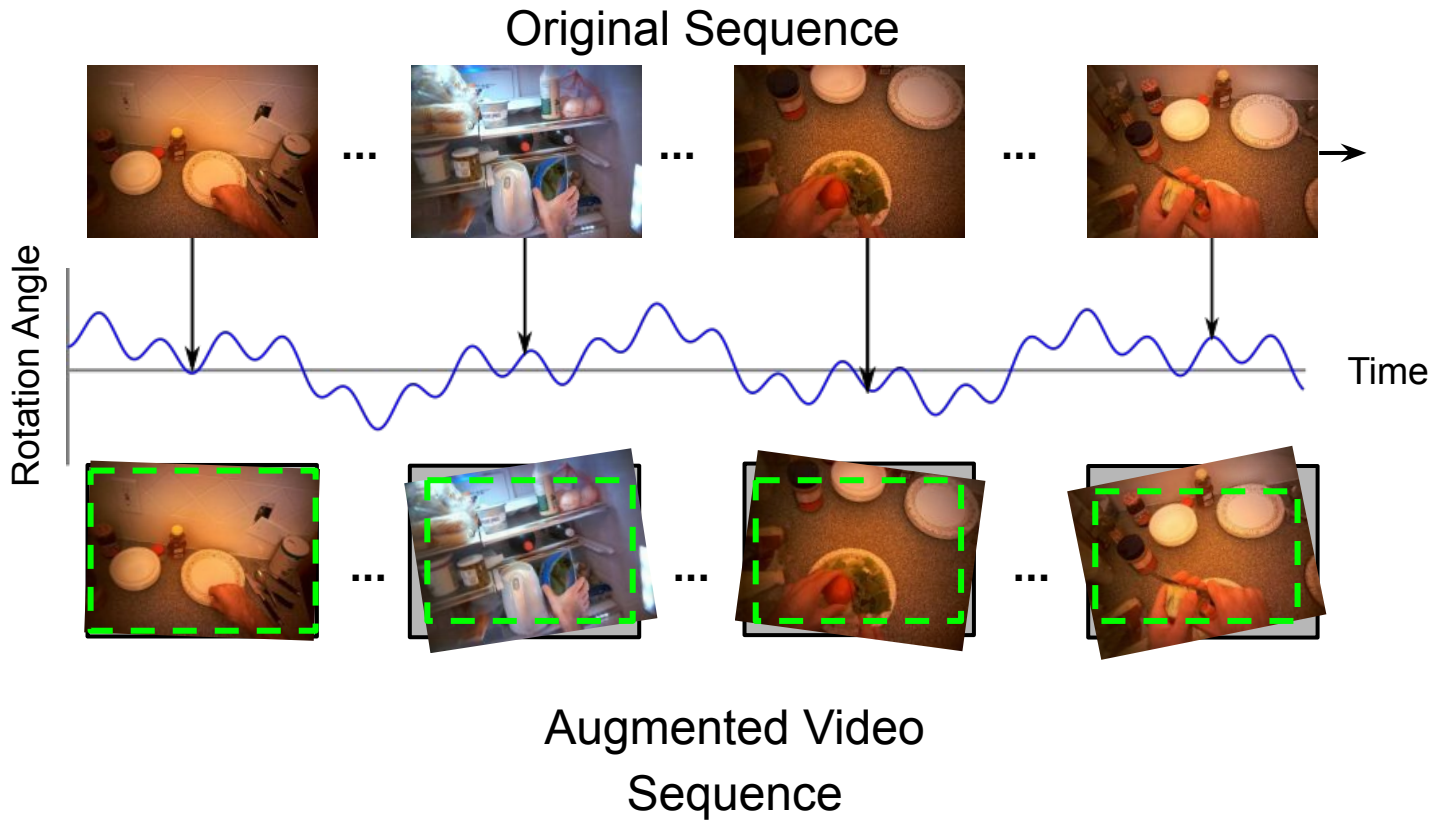


# Methodology | Activity-level Modeling



# Methodology | Visual Data Augmentation

- The purpose is to smoothly rotate the frames.



# Methodology | Sequence Data Augmentation

- Semi-automatically sequence augmentation process.
- A 3-step process:

  1) Cluster **logical action groups**.
  2) Define possible operations: **swap**, **add,** or **remove**.
  3) Randomly apply them.

**Turn on Burner**

**Take Oil**

**Open Oil**

**Pour Oil**

**Close Oil**

**Take mushrooms**

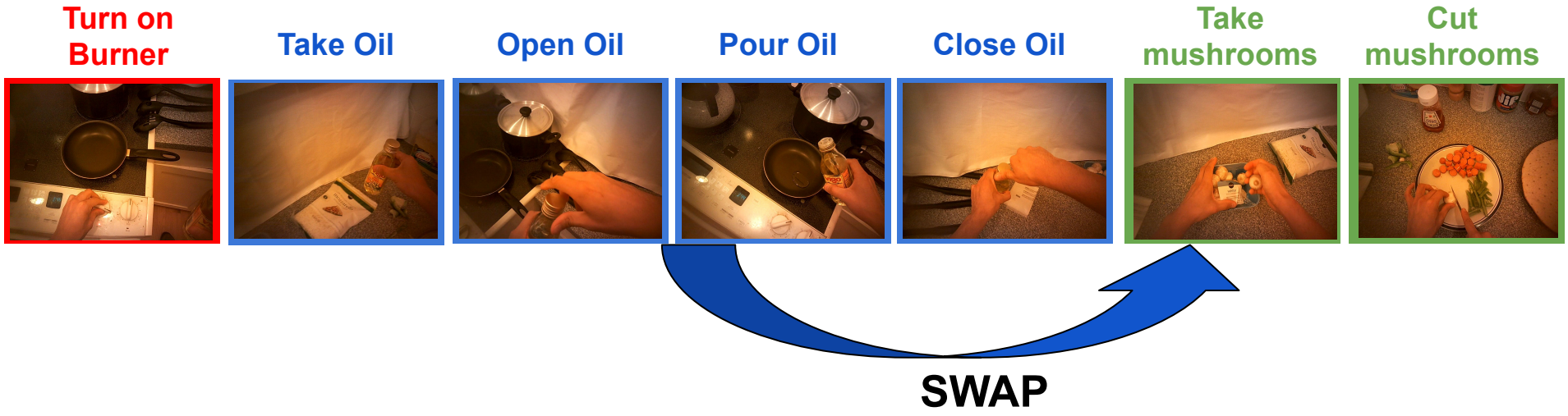
**Cut mushrooms**

# Methodology | Sequence Data Augmentation

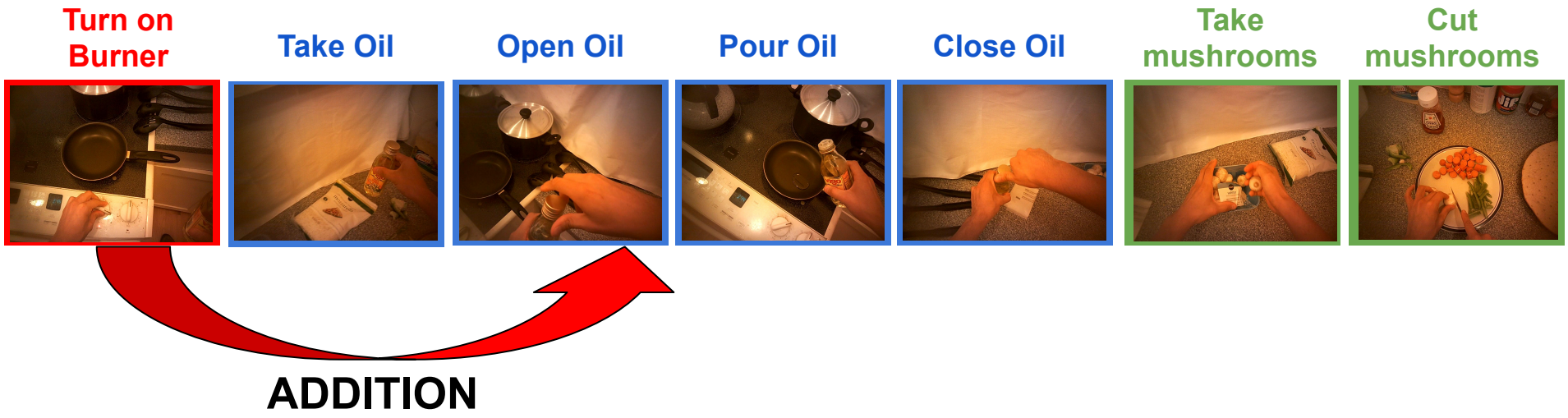
- Semi-automatically sequence augmentation process.
- A 3-step process:

  1) Cluster **logical action groups**.
  2) Define possible operations: **swap**, **add,** or **remove**.
  3) Randomly apply them.

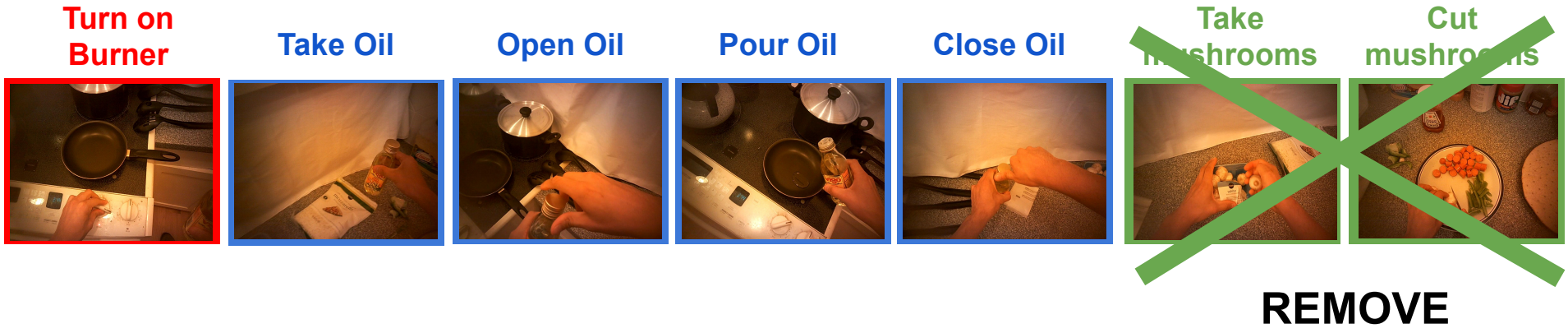
# Methodology | Sequence Operations (1)

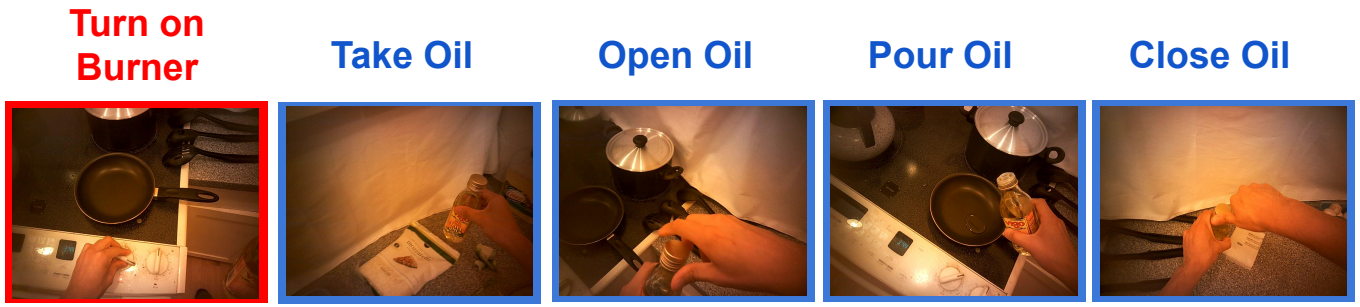


# Methodology | Sequence Operations (2)



# Methodology | Sequence Operations (3)



# Validation | Experimental Datasets

**GTEA**
(Fathi et al., 2011)

**Gaze**
(Fathi et al., 2012)

**Gaze+**
(Li et al., 2015)

| | | | |
|---|---|---|---|
| **# Videos** | 21 | 17 | 37 |
| **# People** | 4 | 14 | 6 |
| **# Actions** | 61 and 71 | 25 and 40 | 40 |
| **# Activities** | 7 | - | 7 |

# Validation | Visual Data Augmentation

## GAZE+ Classification Accuracy

| Shot Evaluation | Sequence | Splits | | | | | | Average |
|---|---|---|---|---|---|---|---|---|
| | | Subject 1 | Subject 2 | Subject 3 | Subject 4 | Subject 5 | Subject 6 | |
| Avg. Prediction | Original | 55.56 | 64.09 | 45.89 | 62.88 | 61.9 | **56.67** | 57.83 |
| | Augmented | **56.55** | **65.77** | **47.58** | **63.16** | **63.27** | 54.17 | **58.41** |
| Weighted Avg. Prediction | Original | 54.56 | **63.76** | 47.79 | 62.88 | **62.59** | **57.08** | **58.11** |
| | Augmented | **56.94** | **63.76** | **49.47** | **63.16** | 60.54 | 54.58 | 58.08 |

# Validation | Sequence Data Augmentation

**GAZE+ Classification Accuracy**

| Shot Evaluation | Sequence | Splits | | | | | | Average |
|---|---|---|---|---|---|---|---|---|
| | | Subject 1 | Subject 2 | Subject 3 | Subject 4 | Subject 5 | Subject 6 | |
| Avg. Prediction | Original | **58.53** | 64.43 | **51.37** | **64.54** | **64.63** | **56.25** | **59.96** |
| | Augmented | 58.33 | **65.44** | **51.37** | 63.99 | **64.63** | 55 | 59.79 |
| Weighted Avg. Prediction | Original | **58.13** | 63.42 | **52.21** | 63.44 | 62.59 | **57.08** | 59.48 |
| | Augmented | 57.74 | **64.09** | 52 | **65.10** | **63.26** | 56.67 | **59.81** |

# Validation | Comparative Results

## Classification Accuracy

| Method | GTEA 61* | GTEA 71 | GAZE 25* | GAZE 40* | GAZE+ | Backbone CNN |
|---|---|---|---|---|---|---|
| CNN Baseline † | 54.67 | 48.95 | 43.44 | 40.76 | 49.83 | VGG-16 Simonyan and Zisserman [2014b] |
| Ours (frame level) | 68.97 | 64.74 | 56.94 | 47.25 | 52.75 | |
| Ours (1 level LSTM) | 69.83 | 71.04 | 63.89 | 49.45 | 58.41 | |
| Ours (Hierarchical LSTM) | 70.69 | 72.95 | **65.28** | **52.75** | 59.96 | |
| Temporal Segments Network Wang et al. [2016] | 67.76 | 67.23 | - | - | 55.25 | ResNet-34 He et al. [2016] |
| LSTA-RGB Sudhakaran et al. [2019b] | 74.14 | 66.16 | - | - | - | |
| Ma et al. [2016] | 75.8 | 73.24 | 62.40 | 43.42 | **66.40** | CNN-M-2048 Chatfield et al. [2014] |
| Sudhakaran and Lanz [2018] | 77.59 | 77 | - | - | 60.13 | ResNet-34 He et al. [2016] |
| LSTA Sudhakaran et al. [2019b] | **79.39** | **78.14** | - | - | - | |

# Conclusions

- Modeling contextual features and their temporal evolution is a promising approach for egocentric activity recognition

- We achieve state of the art results without relying on explicit motion information.