Moving in Sync

Self-supervised Learning of n-Human Interactions

Sonal Sannigrahi

8th October 2021

Outline

Section 1

- Introduction of the problem
- Related Work
- Formalisation

Section 2

- Model Architecture
- Learning Framework

Section 3

- Experiments
- Analysis

Section 4

- Conclusion
- Discussion
- Extension

Outline

Section 1

- Introduction of the problem
- Related Work
- Formalisation

Section 2

- Model Architecture
- Learning Framework

Section 3

- Experiments
- Analysis

Section 4

- Conclusion
- Discussion
- Extension

Method Problem

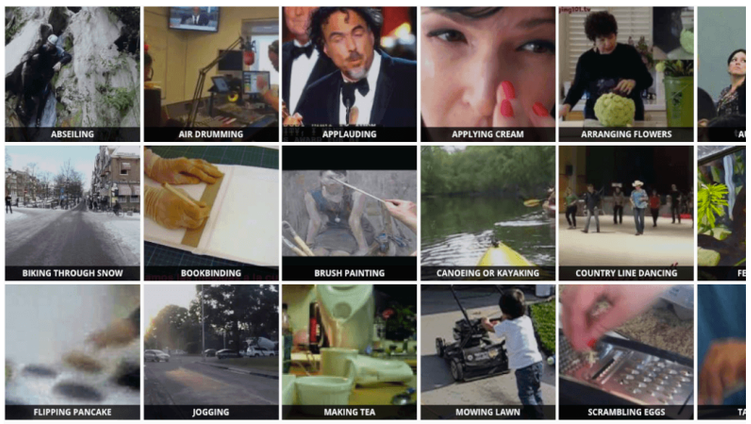
Are they interacting? If so, what are they doing?
How many people are interacting in this scene?

Tasks include:

- Find out what counts as an interaction
- Figure out when it happens
- Retrieve the people who are interacting
- Classify the interaction

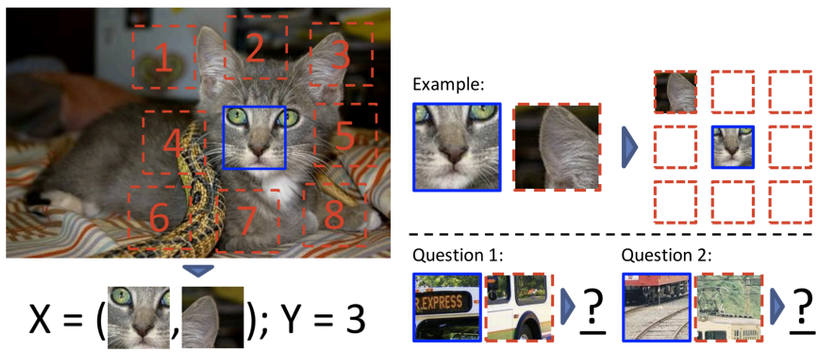
Previous Studies

Action Recognition



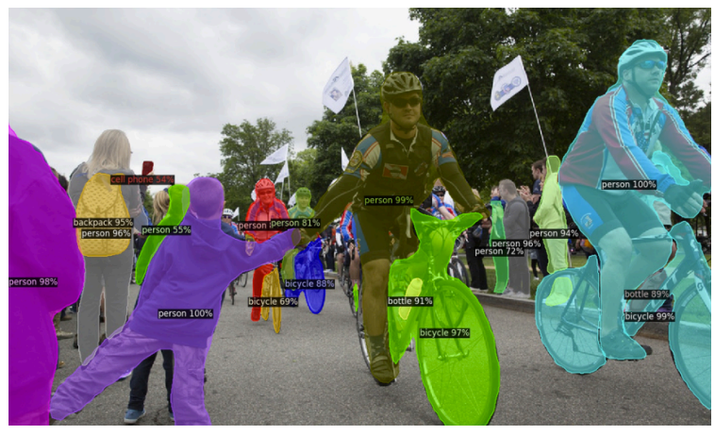
- Convolutional Architectures
- Two-stream networks (RGB+Optical Flow)
- Benefit of Pre-training on Kinetics

Self-supervised Learning



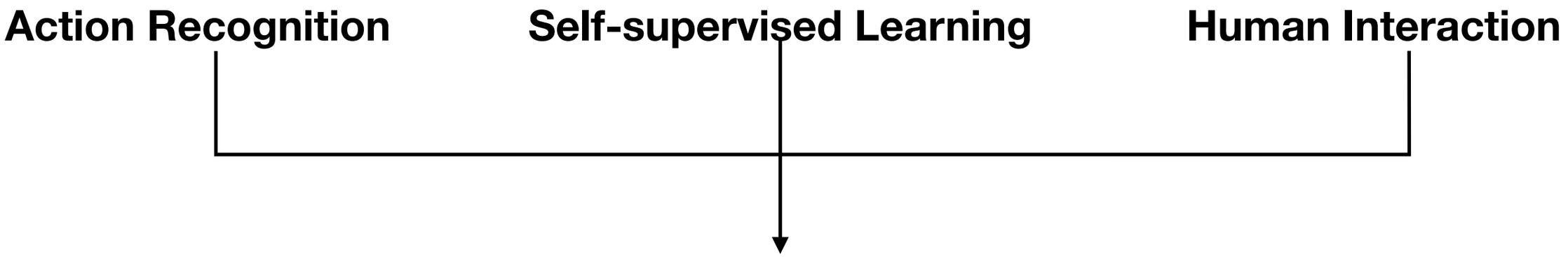
- Convolutional Architectures need large scale data
- Self-supervision in Image and Video domain in popular
- Distinguish positives and negatives via contrastive learning

Human Interaction



- Very little study on Human-Human interactions
- Deep neural networks to extract representations
- …needs well annotated data

Our Approach



Sync Human Tracks in Time and Space via a self-supervised strategy

Outline

Section 1

- Introduction of the problem
- Related Work
- Formalisation

Section 2

- Model Architecture
- Learning Framework

Section 3

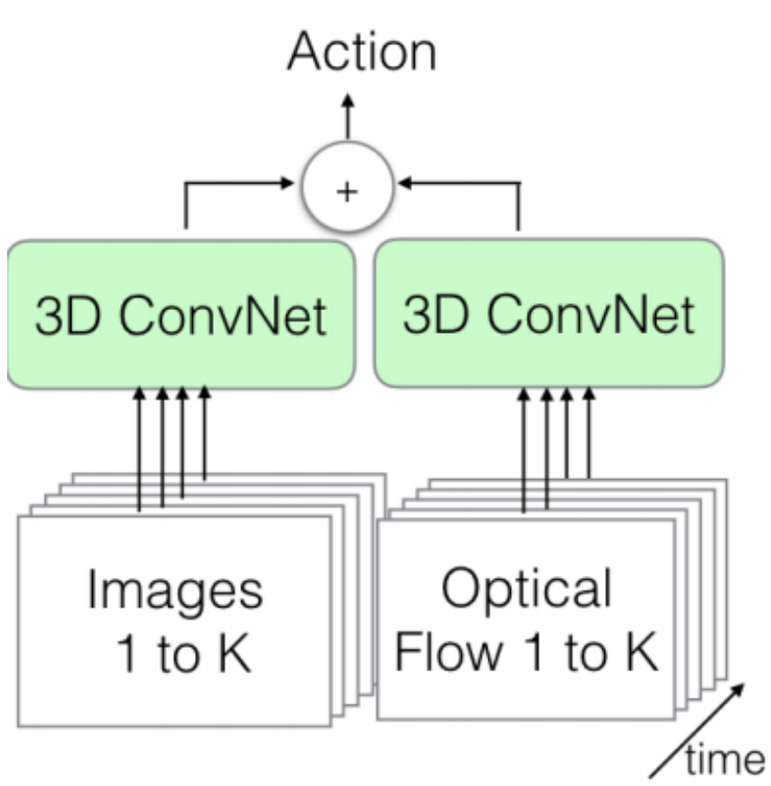
- Experiments
- Analysis

Section 4

- Conclusion
- Discussion
- Extension

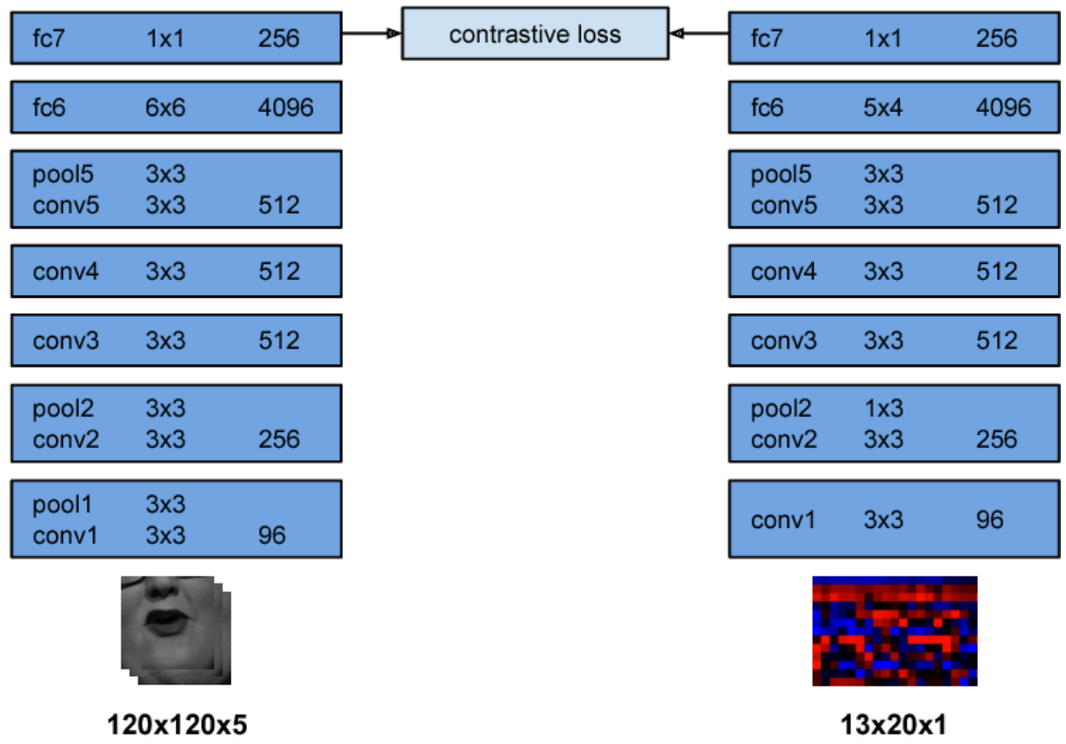
Related Model Architectures

- **I3D**



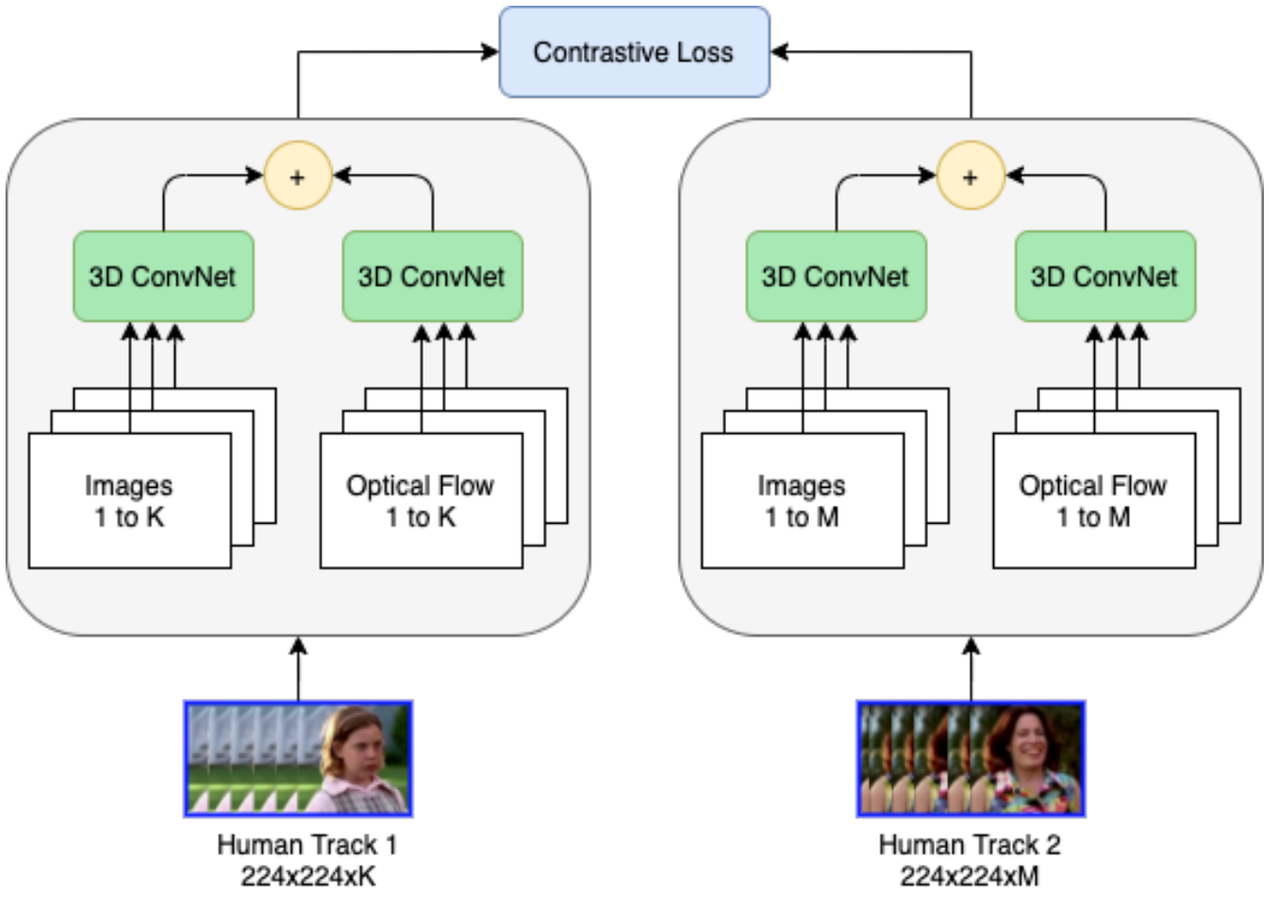
Introduced by J. Carreira and A. Zisserman for the task of **Action Recognition,** CVPR'17

- **SyncNet**



Introduced by JS. Chung and A. Zisserman for the task of **Video/ Audio** Syncing, ACCV'16

Our Proposal



Our Architecture: **Sync3D**

- Siamese two-stream architecture
- Backpropagated via contrastive loss as in SyncNet
- Uses I3D as the main feature extractor
- **Input:** 2 Human Tracks
- **Output:** Interaction label of 1 or 0

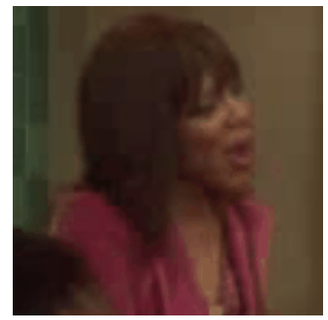
Contrastive Loss

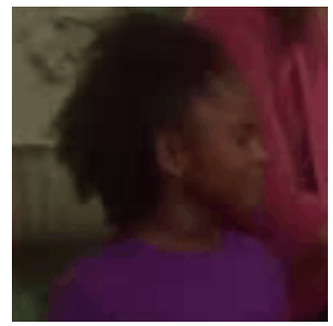
$$L = \frac{1}{2N}\sum_{i=1}^{N} y_n \times d_n^2 + (1 - y_n) max(margin - d_n, 0)^2$$

Ground Truth (1 or 0)

- We “contrast” two feature vectors
- Penalise dissimilar
- Boost similar

Data Sampling

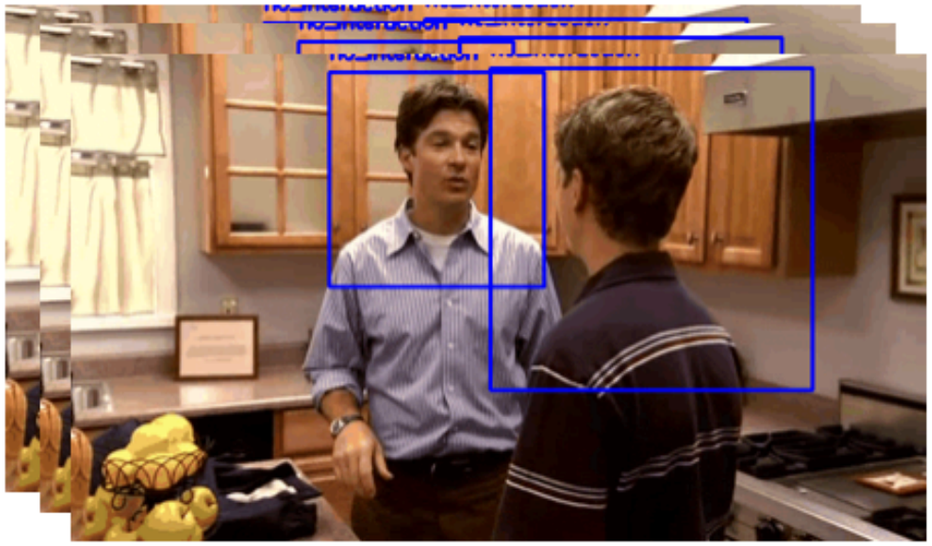




- From an input video, use bounding boxes to get human tracks

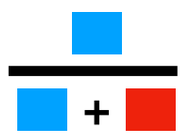
- **Preprocessing Steps:** Crop the frame to the boxes enlarged by 20%, place it on a neutral canvas of 224x224 pixels.

Data Sampling: Positives



Have an Intersection-Over-Union > 0.1

- Get pairs that are temporally and spatially aligned

IOU = 

Happen at the same time

Data Sampling: Negatives

Track 1

Easy Negative

-Take a track from a different video
-This is **class agnostic**

Hard Negative

-Take a track from the same video that is not positive
-This is a **spatial negative**

Medium Negative

-Take a track from the same video that is positive
-Temporally shift by 25 frames
-This is a **temporal negative**

Curriculum Learning

- Replace random sampling of data by a strategic method.
- General sampling: 1/4 Positive and 3/4 Negative
- Our strategy is employed via the addition of hard negatives in different training steps

  Step 100: Start sampling from hard negatives

  Step 200: Increase hard negatives

  Step 300: Add more (1/3 Easy + 1/3 Med + 1/3 Hard)

Outline

Section 1

- Introduction of the problem
- Related Work
- Formalisation

Section 2

- Model Architecture
- Learning Framework

Section 3

- Experiments
- Analysis

Section 4

- Conclusion
- Discussion
- Extension

Dataset: TV-HID[1]

Hand Shake

High Five

Hug

Kiss

- 300 Clips from 23 different TV Shows in 5 different classes: Hug, Kiss, Hand Shake, High Five, Negative
- Labels: bounding box coordinates, interaction, class, head orientation
- **Challenges:** Variety of actors, different camera angles, multiple person scenarios

[1] A. Patron-Perez et al. High Five: Recognising Human Interactions in TV Shows , British Machine Vision Conference, 2010

Evaluation Methodology

- **Self-supervised Task**

Binary classification accuracy for labels 1 and 0

- **Downstream Task of Interaction Classification**

Top1 classification accuracy on the TV-HID labels

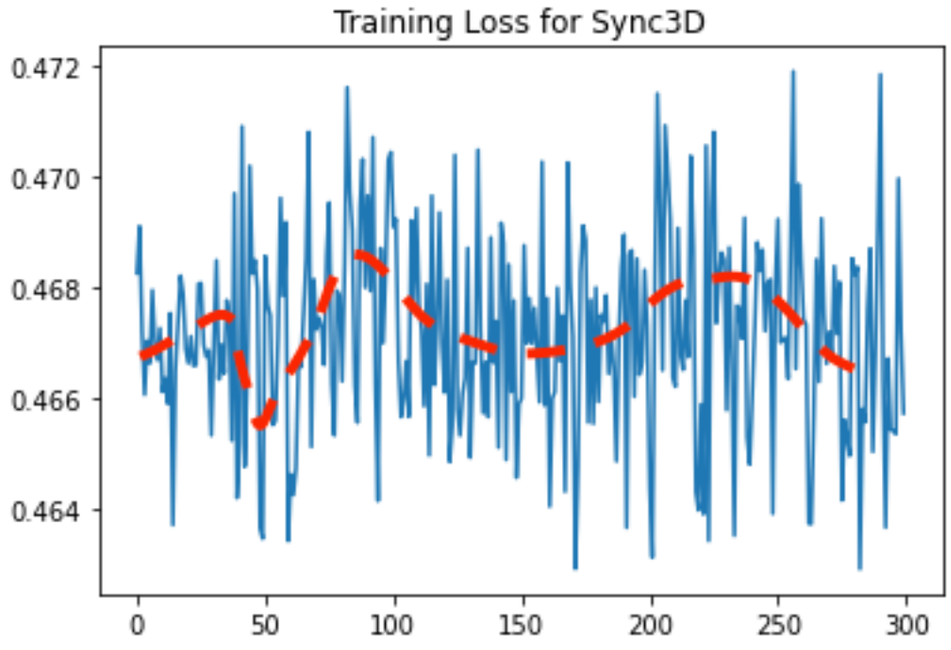
On the self-supervised task

- Test/Train split: 50/50 as provided by TV-HID.
- Sampling Accuracy:

Interaction Class	% Positive	% Negative
Hand Shake	63.45	36.55
High Five	37.25	62.75
Hug	100	0
Kiss	89.10	10.90
Negatives	0	100

TABLE I

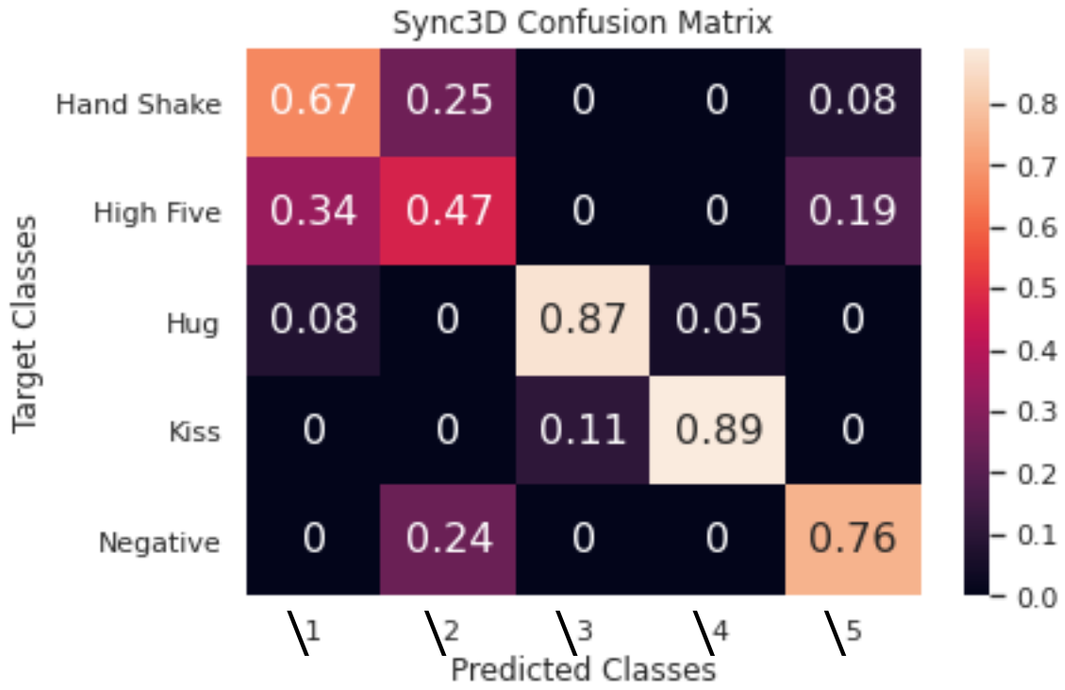
DETECTION ACCURACY FROM OUR SAMPLING STRATEGY

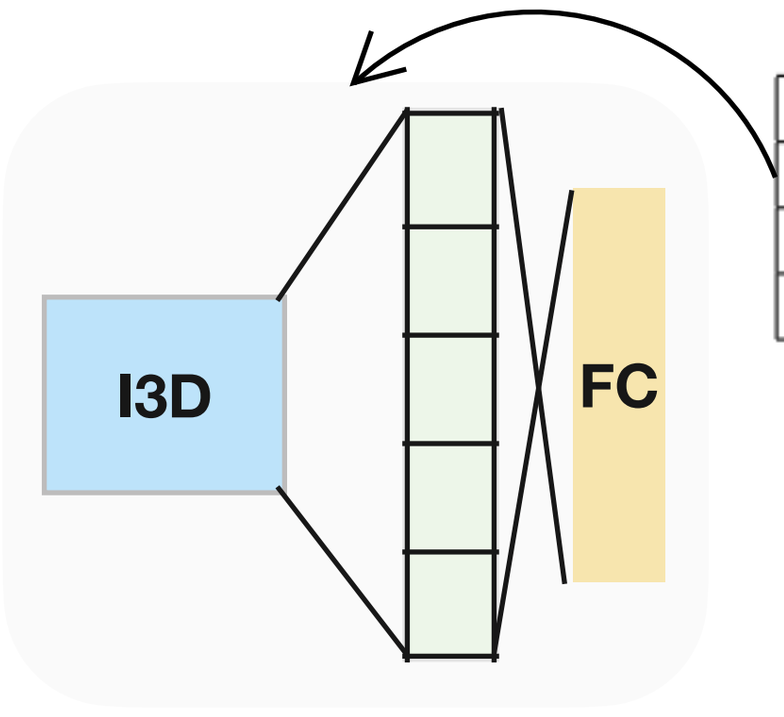


- Training Accuracy: 50.4% on Training Set and 42.6% on Validation Set
- Very poor convergence due to the low number of data pairs.

Downstream Task

- **Good:** Hug, Kiss, Negative
- **Mediocre:** High Five, Hand Shake





Method	setting	curr. learning	% top1 acc.
I3D	Normal	✗	73.2
Sync3D	Easy + Medium Neg	✗	71.3
Sync3D	Hard Neg	✓	75.5

TABLE II

TOP1 CLASSIFICATION ACCURACY FOR I3D AND SYNC3D USING DIFFERENT CURRICULUM LEARNING STRATEGIES

Outline

Section 1

- Introduction of the problem
- Related Work
- Formalisation

Section 2

- Model Architecture
- Learning Framework

Section 3

- Experiments
- Analysis

Section 4

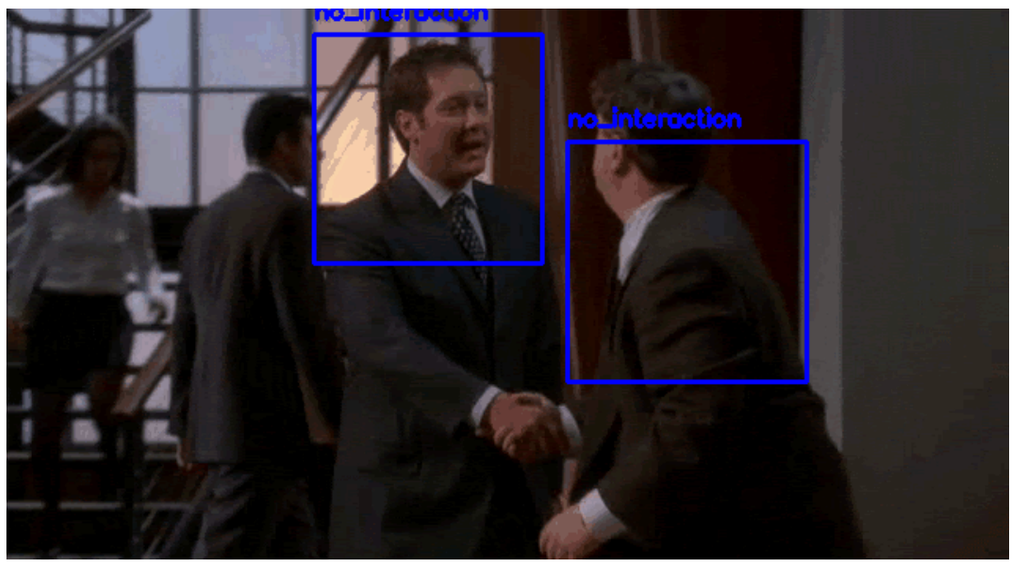
- Conclusion
- Discussion
- Extension

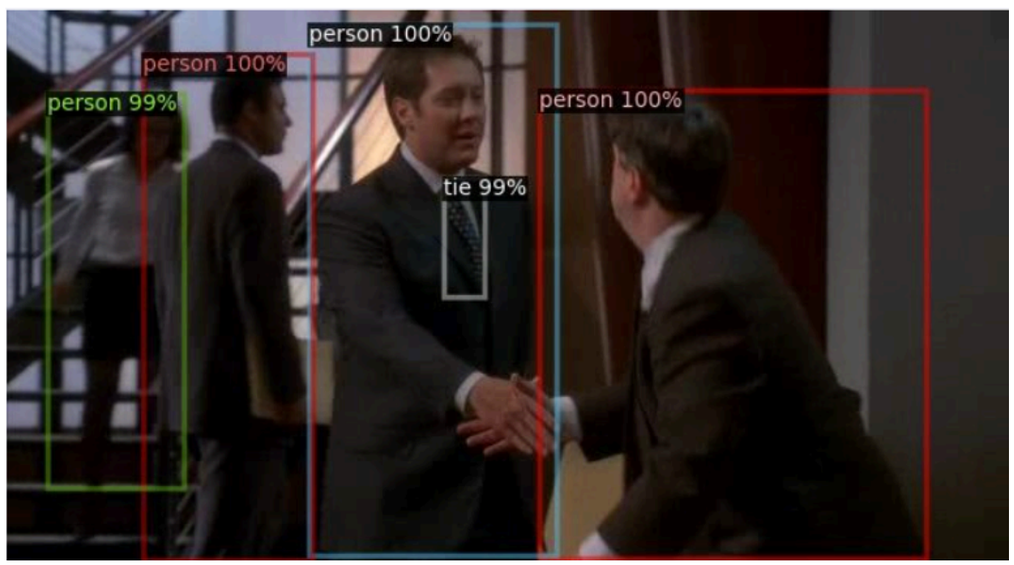
Conclusion

- Introduction of a novel architecture: Sync3D
- Provide a new data sampling strategy for self-supervision
- Evaluation of our feature representation against I3D
- **Future Work:** Flow, Recurrent Architectures

Improvements

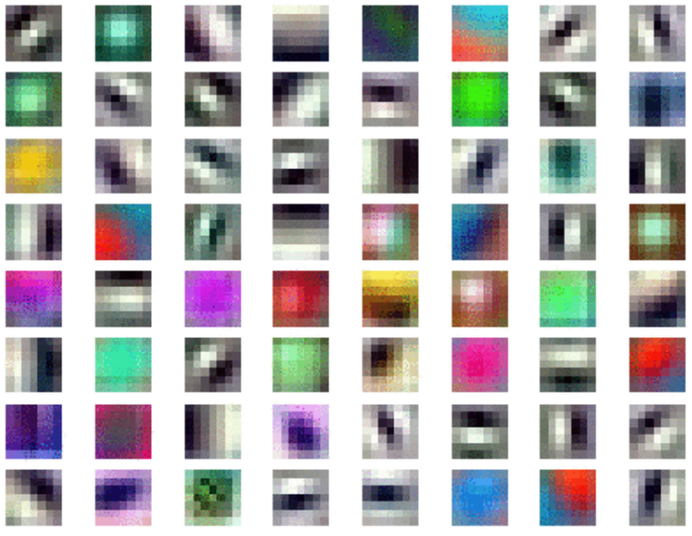
- Train on a larger dataset
- Include a full sliding temporal window
- Use detectron2 detections (full body)
- Further: Head Orientation and Mutual Gaze



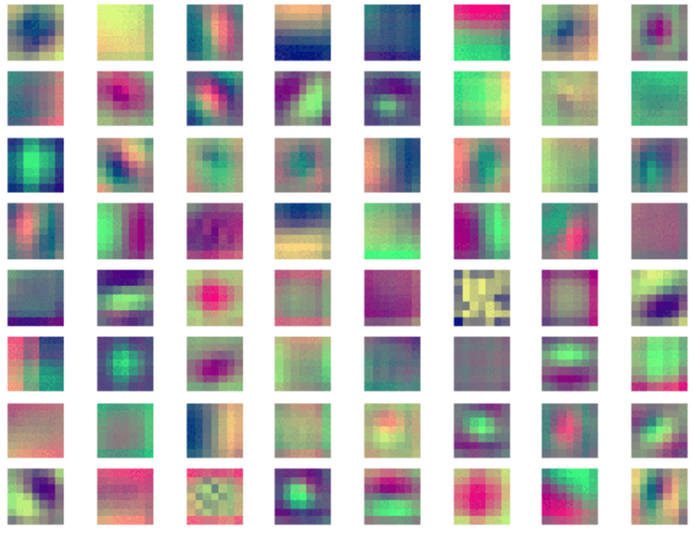


Extensions

- Use Optical Flow
- LSTM-based architectures
- Study of deep convolutional filters



RGB Weights in I3D



Flow Weights in I3D

Thank you!