Temporal Perception and Prediction in Ego-Centric Video

Yipin Zhou Tamara L. Berg

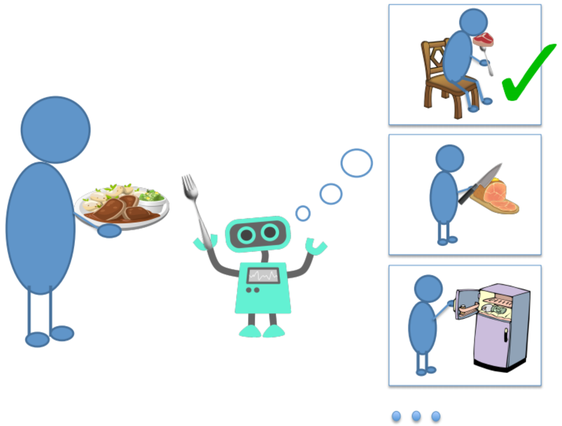
University of North Carolina at Chapel Hill

THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

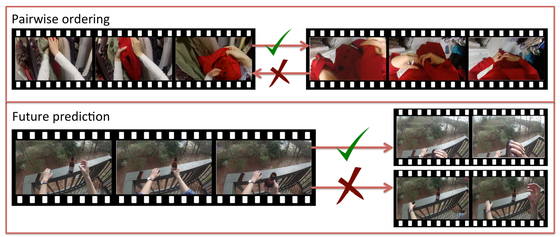
Overview

Abstract:
Given a video of an activity, can we predict what will happen next? In this paper we explore two simple tasks related to temporal prediction in egocentric videos of everyday activities. We provide both human experiments to understand how well people can perform on these tasks and computational models for prediction.

Developing methods for temporal prediction could have far reaching benefits for robots or intelligent agents to anticipate what a person will do, before they do it.



Two tasks:



- In the pairwise ordering task (above) the goal is to provide the correct temporal ordering for two short snippets of video from an activity.
- In the future prediction task (below), given a longer context video of an activity and two video snippets, the goal is to determine which snippet will occur (closest in time) after the context video.

First Person Personalized Activities (FPPA) Dataset

Example frames:

Statistics:

Activities	Avg No.of videos/sub	Avg No.of locs/sub	Totall No.of videos/locs
Wash hands	24.2 (19-34)	3.2 (2-7)	121/16
Put on shoes	22.8 (21-29)	3.0 (2-6)	114/15
Use fridge	26.4 (21-31)	1.6 (1-3)	132/8
Drink water	23.2 (16-31)	3.6 (2-7)	116/18
Put on clothes	21.6 (16-26)	3.4 (2-5)	108/17

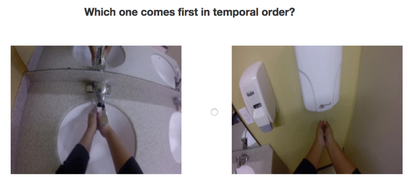
Characteristics:
FPPA dataset enables learning both general and personalized models for temporal prediction.

Human experiments

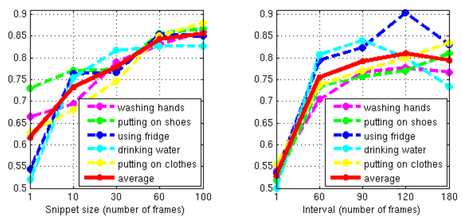
Two experiments:

Snippet size: To evaluate the effect of snippet length on human perceptions of pairwise ordering.

Snippet interval: To explore how the temporal distance between two snippets affects human pairwise ordering performance.



Amazon Mturk Interface

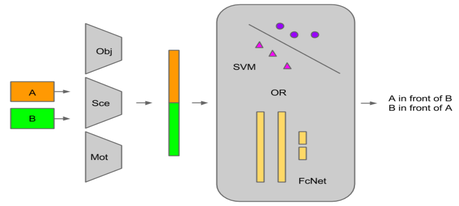


Human performance on pairwise ordering

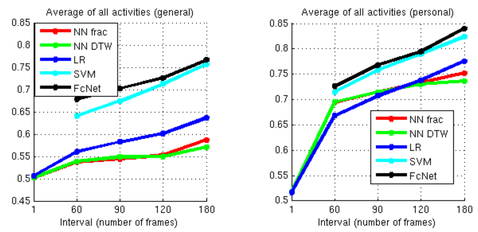
Pairwise ordering task

Video snippet representation:
Object representation + Scene representation +Motion representation

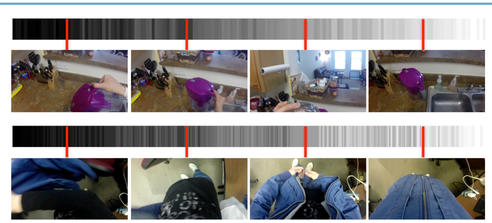
Prediction methods: NN Frac, NN DTW, LR, SVM, FcNet
The first two are nearest neighbor based methods. LR applies linear regression to estimate the temporal position of a video snippet. And the last two directly predict the order of two snippets using linear SVM or a three layer fully-connected network.



Overview of SVM or FcNet method

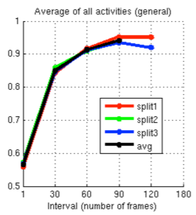


General(left) and personal(right) model performance on pairwise ordering



Inferring temporal information for an entire video sequence

Additional experiments

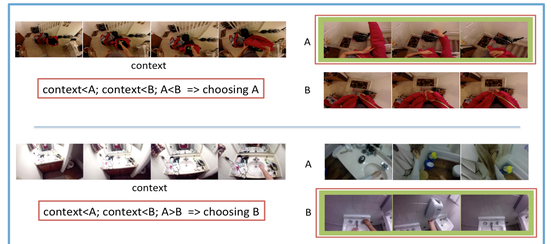


Activities	Our method	FlowWords
wash hands	76.03%	76.86%
put on shoes	69.30%	72.81%
use fridge	79.67%	72.73%
drink water	57.52%	66.09%
put on clothes	80.37%	63.55%
average	72.58%	70.41%

Left is the pairwise ordering accuracy of subset of UCF101 dataset. Right is the forward/backward classification accuracy of our method and Flow-Words method in [1] testing on our dataset.

Future prediction task

Computer prediction:



Visualization results of computer-based future prediction

Human prediction:
We also evaluate how well humans can make future predictions on MTurk.

Activities	SVMg	SVMp	FcNetg	FcNetp	Human
Wash hands	0.6350	0.7550	0.6350	0.7900	0.7816
Put on shoes	0.7000	0.7250	0.7600	0.7700	0.8733
Use fridge	0.6100	0.7100	0.6600	0.7350	0.9284
Drink water	0.6500	0.7300	0.6350	0.7500	0.8717
Put on clothes	0.7100	0.8350	0.6950	0.8650	0.8866
Average	0.6630	0.7510	0.6770	0.7820	0.8686

Future prediction task accuracy by computational methods and people

We also evaluate future prediction results using snippet A and snippet B from different videos. We consider human predictions as ground truth, the general SVM and FcNet models achieve **66.22%** and **66.99%** accuracy respectively.

Supplementary video

References

[1] L. C. Pickup, Z. Pan, D. Wei, Y. Shih, C. Zhang, A. Zisserman, B. Scholkopf, and W. T. Freeman. Seeing the arrow of time. In *CVPR*, 2014.