Developing Motion Code Embedding for Action Recognition in Videos

Maxat Alibayev, David Paulius, and Yu Sun

Department of Computer Science & Engineering, University of South Florida, Tampa, FL, USA.

USFRPAL
Robot Perception and Action Lab

Introduction

- We propose a new embedding of manipulations that encodes actions based on its salient features known as motion codes.
- Such features are mechanical characteristics of manipulation that are relevant to robotics, including contact and trajectory descriptors.
- By including motion codes in a verb classification framework, we can improve their overall performance.
- We demonstrate this improvement on the baseline I3D model.

Motivation

- Humans communicate the idea of manipulation or motion using verbs.
- It can be difficult to convey the meaning of manipulation with words, as human language is inherently ambiguous.
- Conventionally, word vectors have been used to translate words into a machine-interpretable format. Methods such as Word2Vec have been used to derive embeddings.
- However, word vectors derived from models that are learned from text would not capture the true meaning (or distinctions) of motion.
- Motion codes can be used to define representative distances.

Motion Taxonomy

- Motion taxonomy – hierarchical structure of salient features that are relevant to robotics.
- We can build a motion classifier to classify these features from a given input demonstration video.

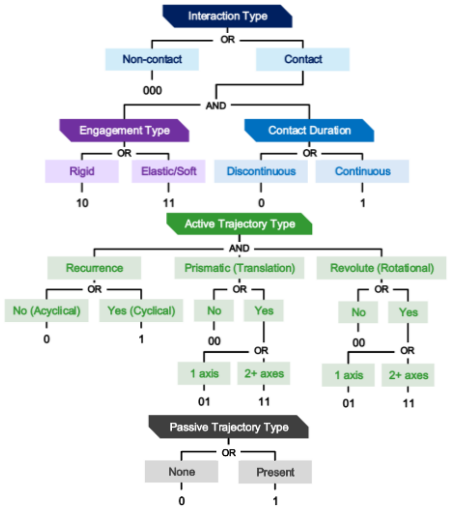

Fig. 1 (from paper) – Motion taxonomy structure

Methodology

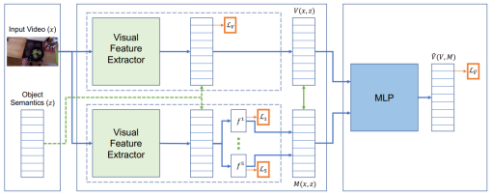

Fig. 2 (from paper) – Verb classification framework

- The verb classification framework combines a probability distribution for verb classes with motion code components into a single feature vector that is passed through a MLP to output a verb prediction.
- Optionally, the framework can use extracted semantic features using objects-in-action used in manipulation.
- Special objective function is defined to train predictors of individual motion code components, which is given as:

$$\mathcal{L}_M = -\sum_{k=1}^{5}\sum_{l=1}^{C_k} \lambda_k m_l^k \log(f_l^k(x))$$

  - f_l^k : – classifier for k-th motion code component
  - λ_k : – constant weight
  - m_l^k : – l-th element of ground-truth vector one-hot vector for k-th motion code component

Implementation

- For visual feature extraction, we used Inflated 3D ConvNet (or I3D) trained on Kinetics dataset [Carreira et al. 2017].
  - Both RGB and optical flow were used.
- Video segments were obtained from EPIC-KITCHENS [Damen et al. 2018] dataset:
  - 3,528 video segments – 2,742 training, 786 validation
  - Sampled 1,517 test videos with 33 verb classes and 32 unique motion codes.
- Training details and parameters:
  - Trained for 50 epochs with Adam optimizer
  - Convolutional layers frozen for first 3 epochs
  - Learning rate set to 0.0003, decreasing by 40% per 5 epochs
  - MLP trained with learning rate of 0.0005 for 200 epochs
  - λ coefficients (from loss) set to 1
  - Video frames sampled to 6 frames / sec
  - 12 consecutive frames randomly sampled, cropped and horizontally flipped as done by NTU-CML-MiRA in EPIC-KITCHENS 2019 challenge.

Experimental Results

- Our objective is to demonstrate the significant improvement obtained once motion codes are integrated in the verb classification pipeline.
- We evaluate the following models:
  - V_x – I3D with visual features only
  - $V_{x,z}$ – I3D with visual features and nouns
  - $\hat{V}(V_x, M_x)$ – our model (I3D + Motion code embedding)
  - $\hat{V}(V_x, M_{x,z})$ – our model with nouns (I3D + Motion code embedding)
- Our testing results show that motion code prediction improves verb classification as compared to baseline model.

TABLE III
VERB CLASSIFICATION COMPARISON ON TESTING VIDEOS BASED ON TOP-1 ACCURACY (AS %).

Methods	RGB	Flow	Fused
Baseline, V_x	33.36	31.64	36.12
Motions, $\hat{V}(V_x, M_x)$	33.62	32.30	36.78
Motions with nouns, $\hat{V}(V_x, M_{x,z})$	**34.08**	**34.74**	**38.04**
Baseline with nouns, $V_{x,z}$	38.69	38.76	41.73
Motions, $\hat{V}(V_{x,z}, M_x)$	**38.89**	37.05	42.06
Motions with nouns, $\hat{V}(V_{x,z}, M_{x,z})$	38.83	**39.95**	**42.12**

- Our validation results also show that verb classification can be improved, granted that motion code prediction is accurate.

TABLE IV
VERB CLASSIFICATION COMPARISON ON VALIDATION VIDEOS BASED ON TOP-1 ACCURACY (AS %).

Methods	RGB	Flow	Fused
Baseline, V_x	41.60	39.82	45.04
Baseline with nouns, $V_{x,z}$	48.22	44.15	49.24
Predicted Motions, $\hat{V}(V_x, M_x)$	41.22	40.46	46.18
Predicted Motions with nouns, $\hat{V}(V_x, M_{x,z})$	43.13	42.11	47.20
Ground Truth Motions*, $\hat{V}(V_x, M_x)$ (Using true motion code as embedding)	**53.82**	**53.69**	**57.63**

Acknowledgement

This material is based upon work supported by the National Science Foundation under Grant Nos. 1812933 and 1910040.