Spatio-Temporal Naive Bayes Nearest Neighbor (ST-NBNN) for Skeleton-based Action Recognition

Junwu Weng *Chaoqun Weng* *Junsong Yuan*

School of EEE, Nanyang Techonological University, Singapore

IEEE 2017 Conference on Computer Vision and Pattern Recognition

CVPR July 21-26 HONOLULU 2017

Introduction

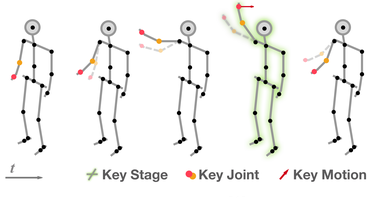
Problem Statement

Input : pre-segmented 3D action sequence

Output : action label

Label : Waving Right Hand

Motivation

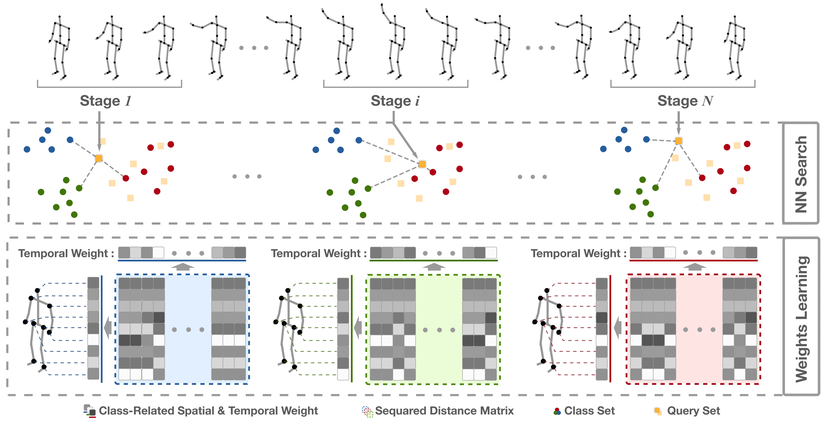


- Non-parametric model like NBNN[1] has not been well explored before in skeleton-based action recognition task
  - > NBNN was successful in image classification
  - > 3D action recognition is a light-weight problem
- The recognition of a certain action only relates to the movement of a subset of joints (spatially) and a few certain frames (temporally)

Overview

- 3D action sequences are separated into pre-defined number of stages
- NBNN is used as the main framework for action classification
- Spatio-temporal weights are involved to help improve recogntion performance and discover key spatio-temporal patterns at the same time

Proposed Method



Formulation

NBNN : $c^* = \arg\min_c \sum_{i=1}^{N} \|\boldsymbol{x}^i - NN_c(\boldsymbol{x}^i)\|^2 = \arg\min_c \mathbf{1}^\intercal \boldsymbol{X}_c \mathbf{1}$

ST-NBNN : $c^* = \arg\min_c (\boldsymbol{u}_c^s)^\intercal \boldsymbol{X}_c \boldsymbol{u}_c^t$ (Bi-linear Function[2])

Weight Learning

- Fix $\boldsymbol{u}_c^t$ and Update $\boldsymbol{u}_c^s$

$$\min_{\boldsymbol{u}_c^s} \frac{1}{2}\beta_1 \|\boldsymbol{u}_c^s\|^2 + \lambda \sum_{i=1}^{K} max(0, 1 - c_i f_c(\boldsymbol{X}_c^i))^2$$

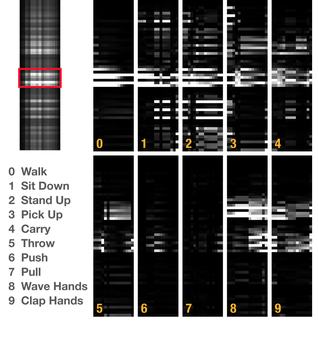
$$\beta_1 = \|\boldsymbol{u}_c^t\|^2$$

- Fix $\boldsymbol{u}_c^s$ and Update $\boldsymbol{u}_c^t$

$$\min_{\boldsymbol{u}_c^t} \frac{1}{2}\beta_2 \|\boldsymbol{u}_c^t\|^2 + \lambda \sum_{i=1}^{K} max(0, 1 - c_i f_c(\boldsymbol{X}_c^i))^2$$

$$s.t \quad \sum_{i=1}^{N} \boldsymbol{u}_c^t(i) = N, \quad \boldsymbol{u}_c^t \succeq \mathbf{0}$$

$$\beta_2 = \|\boldsymbol{u}_c^s\|^2$$

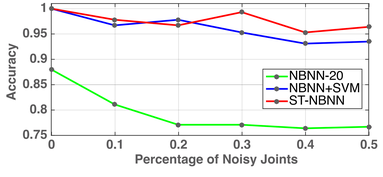


Results

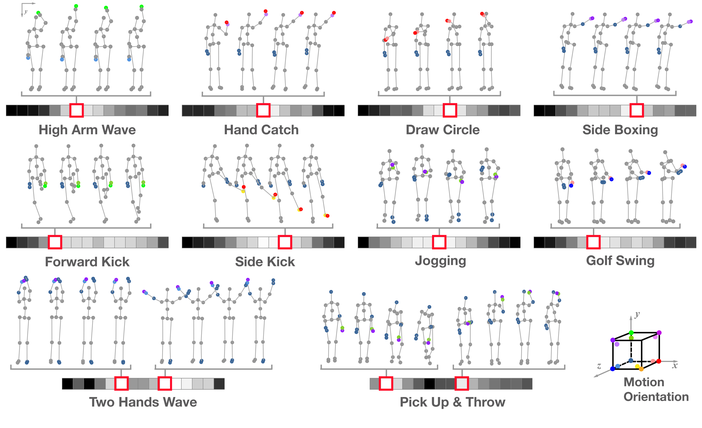
Recognition Accuracy

Method	MSR	UTK	UCB
NBNN-N	91.7	95.5	88.0
NBNN+SVM	92.4	94.0	100.0
Best Method	94.8[3]	98.2[4]	100.0[5]
Ours	*94.8*	*98.0*	*100.0*

Robustness Analysis



Discovered Key Joints, Stages and Motion



Reference

[1] O. Boiman, E. Shechtman, and M. Irani. In defense of nearest-neighbor based image classification. In CVPR'08

[2] H.Pirsiavash,D.Ramanan,andC.C.Fowlkes.Bilinearclas- sifiers for visual recognition. In NIPS'09

[3] P. Wang, C. Yuan, W. Hu, B. Li, and Y. Zhang. Graph based skeleton motion representation and similarity measurement for action recognition. In ECCV'16

[4] P. Koniusz, A. Cherian, and F. Porikli. Tensor representations via kernel linearization for action recognition from 3d skeletons. In ECCV'16

[5] J. Liu, A. Shahroudy, D. Xu, and G. Wang. Spatio-temporal lstm with trust gates for 3d human action recognition. In ECCV'16

Acknowledgement

This work is supported in part by Singapore Ministry of Education Academic Research Fund Tier 2 MOE2015-T2-2-114.