KEY JOINTS SELECTION AND SPATIOTEMPORAL MINING FOR SKELETON-BASEDACTION RECOGNITION

Zhikai Wang,Chongyang Zhang,Wu Luo, and Weiyao Lin
School of Electronic Information and Electrical Engineering,Shanghai Jiao Tong University, Shanghai 200240, China
Shanghai Key Lab of Digital Media Processing and Transmission, Shanghai 200240, China

Introduction

- The histograms of human joints positions and velocitys are developed in order to enhance the spatiotemporal structure representation.
- The key joints are selected based on their information gain, then the histograms are weighted and composed with trajectory features.

Main Contributions:

- Introduce a novel action recognition framework using Key Joints Selection and Spatiotemporal Mining, which can identify both key joints and their position & velocity histogram as well as trajectory features for action classification.

Pipeline of the proposed framework

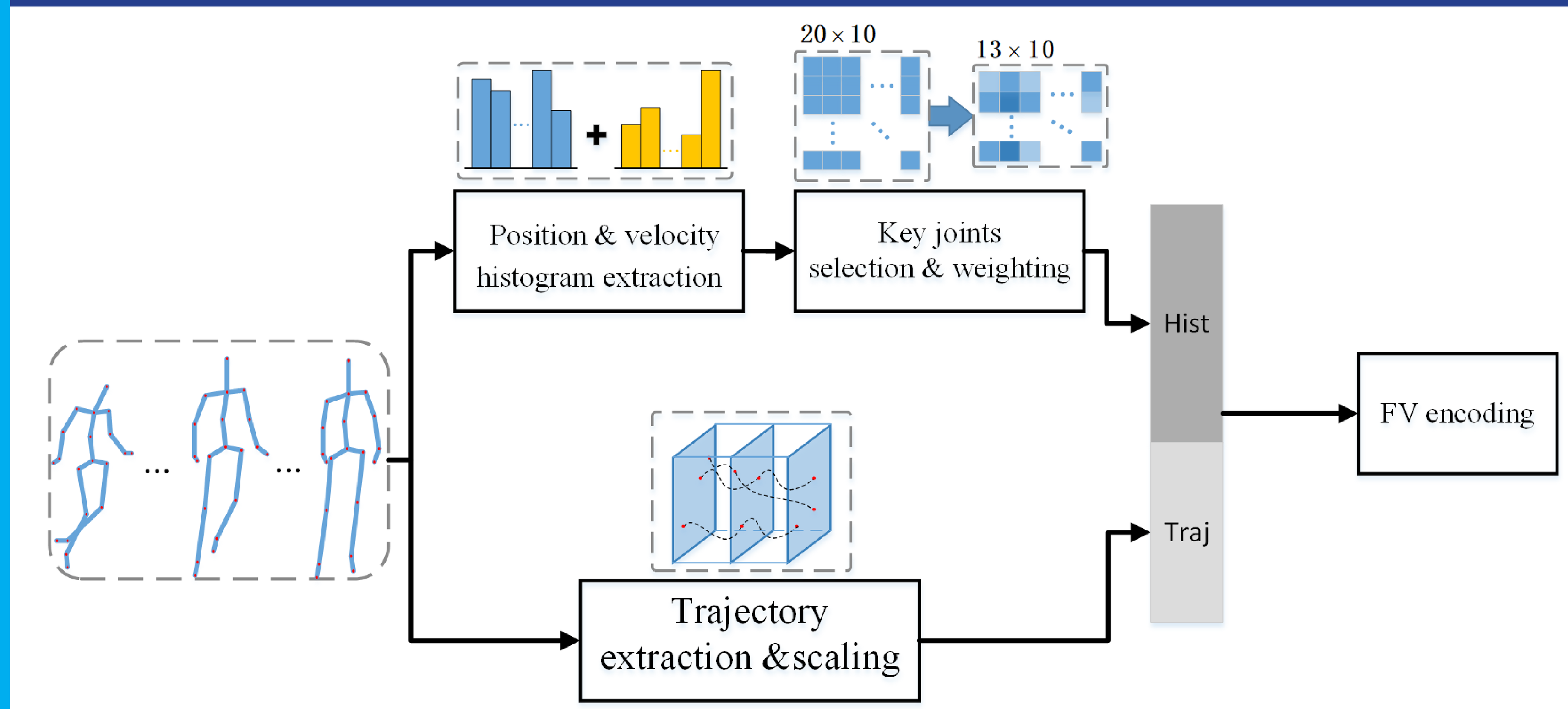

Fig. 1. The pipeline of the proposed framework. Firstly, two features are extracted and processed from skeleton-based action sequences:1) position & velocity histogram extraction and then key joints selection & weighting, and 2) trajectory extraction and scaling . Secondly, the two type features are concatenated and then encoded into fisher vector as final feature for classification.

Position & velocity histogram construction(1)

As shown in Fig. 2, each skeleton-based action sampleis a sequence of 3D pose frames. For one joint j (Totally J joints in all) in a sample with F frames, its 3D temporal position and velocity features are formulated as the following:

$$p_j = [p_{j,1}^T, \dots p_{j,F}^T]^T \in R^{3F}$$

$$v_j = [v_{j,1}^T, \dots v_{j,F}^T]^T \in R^{3F}$$

Correspoding Author: Chongyang Zhang
E-mail : sunny_zhang@sjtu.edu.cn

Position & velocity histogram construction(2)

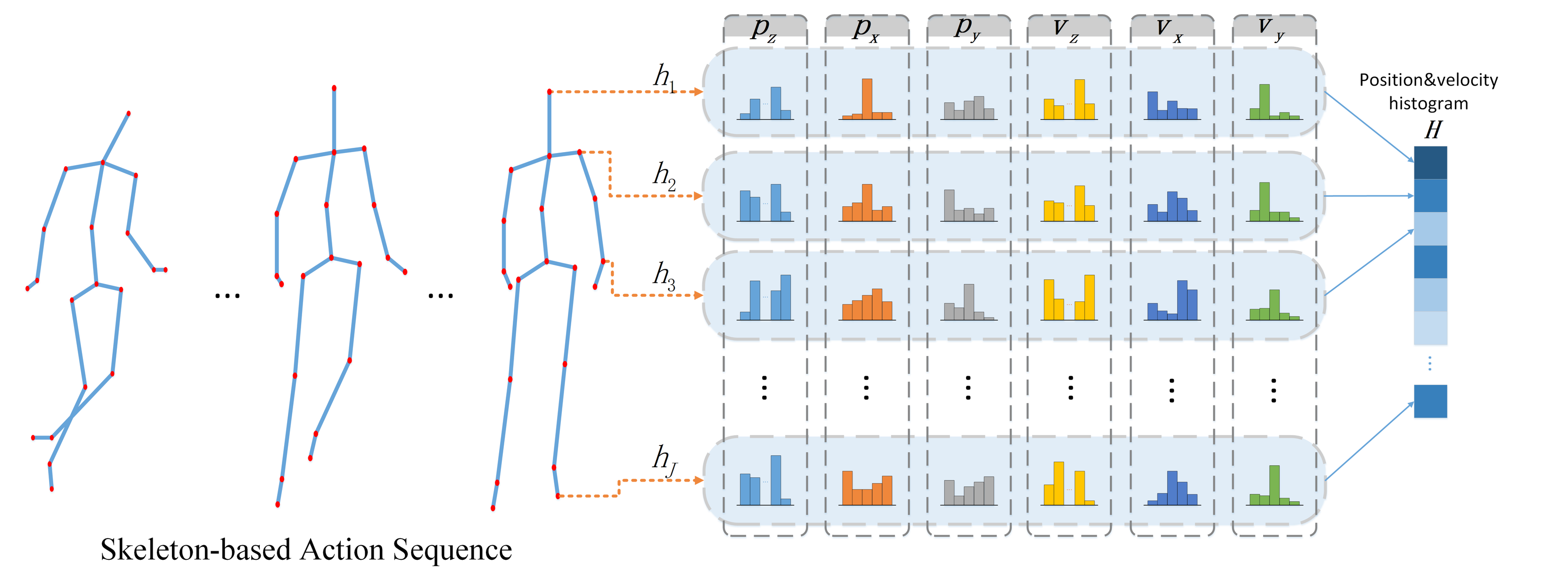

Fig. 2. The construction of histogram feature. For each joint, there are six dimensions of histograms: the x,y,z dimensions of position and velocity, respectively. The six histograms for every joints are then concatenated to form one full position & velocity histogram for each action sequence.

The joint j's descriptor is denoted as: $x_j = [p_j^T, v_j^T]$, and each component of pj and vj, has three channels to denote the value of x,y,z dimensions of position and velocity, respectively. The six parts are mapped into six histograms for each joint. In Fig. 3, they are denoted as Pz; Px; Py; Vz; Vx; Vy; respectively, and the concatenation of them forms the histogram hj of xj. All the hj of J joints are then concatenated as $H = \{h_j\}_{j=1}^J$ to describe the joints' position and velocity distribution in one action sequence.

Key joints selection and weighting

- **Selection:** In most actions, only a few joints are responsible for the action recognition, so we propose to only preserve the most informative joints for classification based on infomation gain.

$$Ent(D,T) = -\sum_{k=1}^{n} p_k \log_2(p_k)(p_k = \frac{TP_k}{m_k})$$

Here n is the number of categories, TPk denotes the number of true positives for class k, mk means the number of samples of class k, and pk is to present the test data's "purity". The information gain by h_j can be obtained by:

$$Gain(D,T,h_j) = Ent(D,T) - Ent(D,T,h_j)$$

here $Ent(D,T,h_j)$ is the reclaculation of entropy after joint j's histogram used. The joints with highest IG will be selected.

- **Weighting:** Spatiotemporal weighting can better leverage the unequal contribution of different joints in different stages for action recognition. We use the spatiotemporal relative mean velocity to measure the weight of a joint's certain stage.

$$\bar{v}_{j,s} = \frac{1}{N}\sum_{f=1}^{N} v_{j,s,f}$$

Here:
is the mean velocity of joint j in stage s with N frames.

Quantitative Results

Table 1. Comparison of Results on MSRAction3D (%)

method	AS1	AS2	AS3	Average
Lie Group [1]	**0.954**	0.839	**0.982**	0.925
SCK+DCK [2]	-	-	-	0.94
HBRNN [3]	0.933	0.946	0.955	0.945
ST-LSTM [4]	-	-	-	0.948
GRAPH-Based [5]	0.936	0.955	0.951	0.948
ST-NBNN [6]	0.915	0.956	0.973	0.948
Ours	0.945	**0.96**	0.973	**0.959**

Discussion

- Key joints selection and spatiotemporal weighting is well discover critical patterns for skeleton-based action recognition.
- Compared to only trajectories-based method, the introduction of histograms enhances the spatial representation.("wave" vs "high wave")
- The introduction of velocity histograms improve the orientation discrimination. ("pull" vs "push")
- We can give one reasonable explain why the proposed method perform better in MSRAction3D is that: the samples in MSRAction3D have more inter-class difference in key joints' position changes, and thus the proposed framework, which combined position histogram and trajectory features, can get higher recognition accuracy.

Reference

[1] R. Vemulapalli, F. Arrate, and R. Chellappa, Human action recognition by representing 3d skeletons as points in a lie group.In CVPR, 2014.
[2] P. Koniusz, A. Cherian, and F. Porikli, Representations via kernel linearization for action recognition from 3d. In arXiv, 2016.
[3] Y. Du, W. Wang, and L. Wang, Hierarchical recurrent neural network for skeleton based action recognition. In CVPR, 2015.
[4] J. Liu, A. Shahroudy, D. Xu, and G. Wang, Spatiotemporal lstm with trust gates for 3d human action recognition. In ECCV, 2016.
[5] P. Wang, C. Yuan, W. Hu, B. Li, and Y. Zhang, Graph based skeleton motion representation and similarity measurement for action recognition. In ECCV, 2016,
[6] J. Weng, C. Weng, and J. Yuan, Spatio-temporal naivebayes nearest-neighborfor skeleton-based action recognition. In CVPR, 2017.