Pattern Recognition 55 (2016) 148–159

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

A novel hierarchical framework for human action recognition

Hongzhao Chen [a], Guijin Wang [a,*], Jing-Hao Xue [b], Li He [a]

[a] Department of Electronic Engineering, Tsinghua University, Beijing 100084, China
[b] Department of Statistical Science, University College London, London WC1E 6BT, UK

ARTICLE INFO

Article history:
Received 3 July 2015
Received in revised form
15 January 2016
Accepted 16 January 2016
Available online 30 January 2016

Keywords:
Action recognition
3D skeleton
Hierarchical framework
Part-based
Time scale
Action graphs

ABSTRACT

In this paper, we propose a novel two-level hierarchical framework for three-dimensional (3D) skeleton-based action recognition, in order to tackle the challenges of high intra-class variance, movement speed variability and high computational costs of action recognition. In the first level, a new part-based clustering module is proposed. In this module, we introduce a part-based five-dimensional (5D) feature vector to explore the most relevant joints of body parts in each action sequence, upon which action sequences are automatically clustered and the high intra-class variance is mitigated. In the second level, there are two modules, motion feature extraction and action graphs. In the module of motion feature extraction, we utilize the cluster-relevant joints only and present a new statistical principle to decide the time scale of motion features, to reduce computational costs and adapt to variable movement speed. In the action graphs module, we exploit these 3D skeleton-based motion features to build action graphs, and devise a new score function based on maximum-likelihood estimation for action graph-based recognition. Experiments on the Microsoft Research Action3D dataset and the University of Texas Kinect Action dataset demonstrate that our method is superior or at least comparable to other state-of-the-art methods, achieving 95.56% recognition rate on the former dataset and 95.96% on the latter one.

1. Introduction

Action recognition is an active research topic that focuses on labeling a motion sequence as one of the known actions. It can be widely applied in human–computer interaction, health care, video surveillance, etc. In order to achieve high accuracy and great robustness for real-world applications, an action recognition system has to overcome three challenges: high intra-class variance with low inter-class variance, variable movement speed, and high computational costs. As shown in Fig. 1, people may perform the same action of *Side Boxing* in quite different ways, by using one hand or two hands, leading to high intra-class variance. Meanwhile, people may also perform the same action with variable movement speed, as demonstrated in Fig. 2.

Prior to 2010, many color image-based methods of action classification had been studied [1]. However, these methods have relatively low recognition accuracies and thus they are unable to be applied in real-world applications.

The situation has been much improved as technologies on depth imaging advance quickly [2–5]. Recent works of action recognition could be divided into two types, depth map-based methods [6–10] and 3D skeleton-based methods [11–22]. The former directly takes sequences of depth maps as input, while the latter utilizes 3D skeleton sequences inferred from depth maps. Fig. 3 shows the color image, depth image and 3D skeleton acquired from a Kinect sensor.

Depth map-based methods extract features from depth maps to describe the human poses and model the transition of poses. The widely-used features include sampled 3D points from silhouettes [6,7], histograms of oriented gradients [8], histogram of oriented 4D normals [9], histogram of oriented principal components [10], etc. However, the extraction of these features is often time-consuming, making them hardly applicable in real-time scenarios.

In fact 3D human skeleton, which could be reliably estimated from depth maps in real time [23–25], is an efficient and concise surrogate to describe the human poses. In most 3D skeleton-based methods [11,12,14,16,18,21], motion features are represented by pair-wise differences of joint positions within the current frame or between the current frame and the previous frames. Hence motion features extracted from 3D skeleton can efficiently model the action dynamics. Kapsouras et al. [21] further considered the time scale for motion features to fit various movement speeds. However, no principle has been supplied yet for how to determine the time scale. Some other histogram-based features are also proposed, like histograms of 3D joints [13], space time pose [17], histogram of oriented displacements [15], points in a Lie group [19], etc.

* Corresponding author. Tel.: +86 18911389502; fax: +86 62770317.
E-mail addresses: jordanchan1004@163.com (H. Chen), wangguijin@tsinghua.edu.cn (G. Wang), jinghao.xue@ucl.ac.uk (J.-H. Xue), hhappy06@gmail.com (L. He).

http://dx.doi.org/10.1016/j.patcog.2016.01.020
0031-3203/

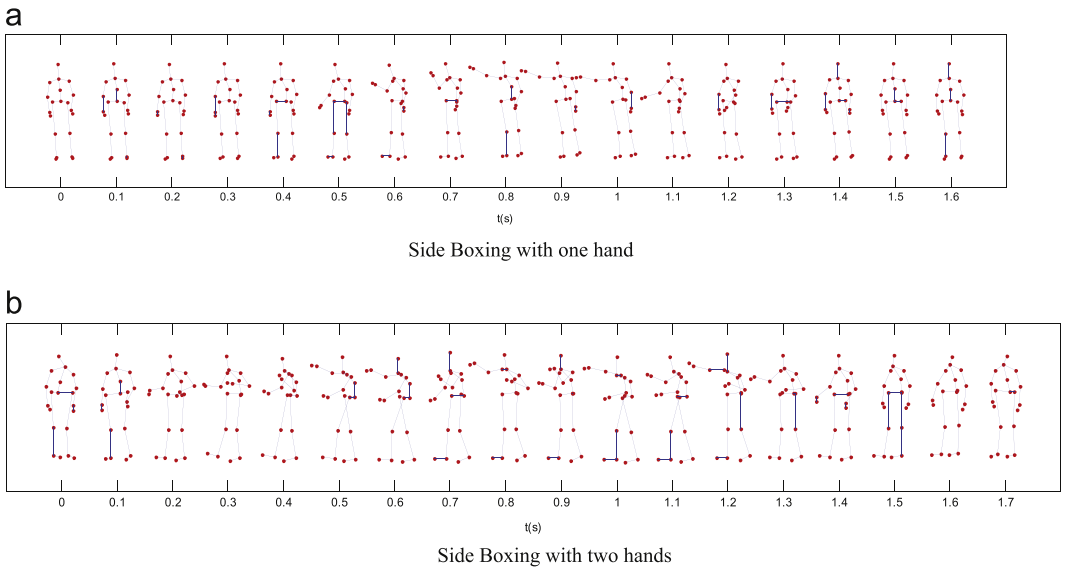

Fig. 1. An illustrative example of high intra-class variance. Two panels present skeleton sequence diagrams of action *Side Boxing* sampled at 10 fps.

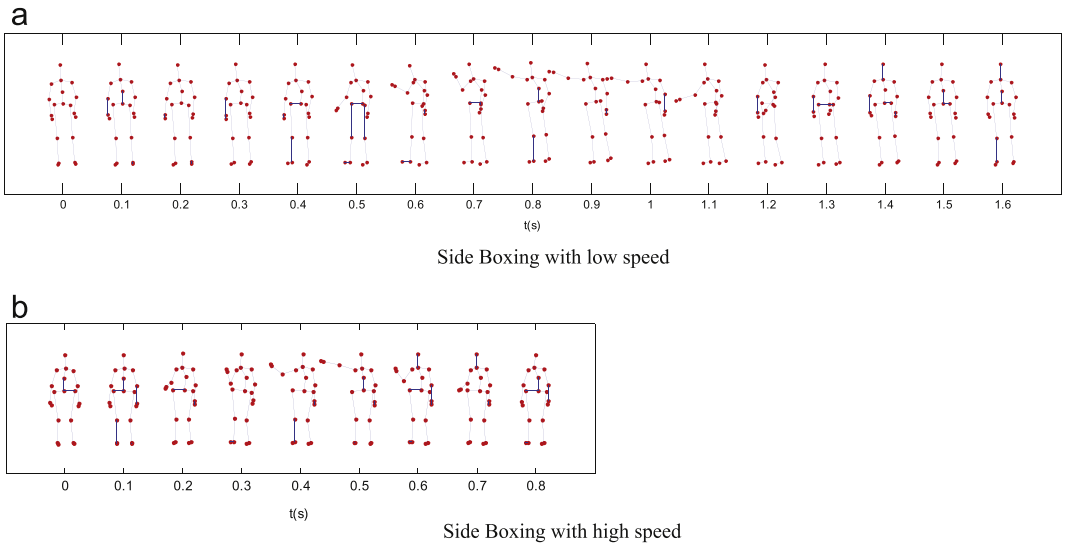

Fig. 2. An illustrative examples of variable movement speed. Two panels present skeleton sequence diagrams of action *Side Boxing* sampled at 10 fps.

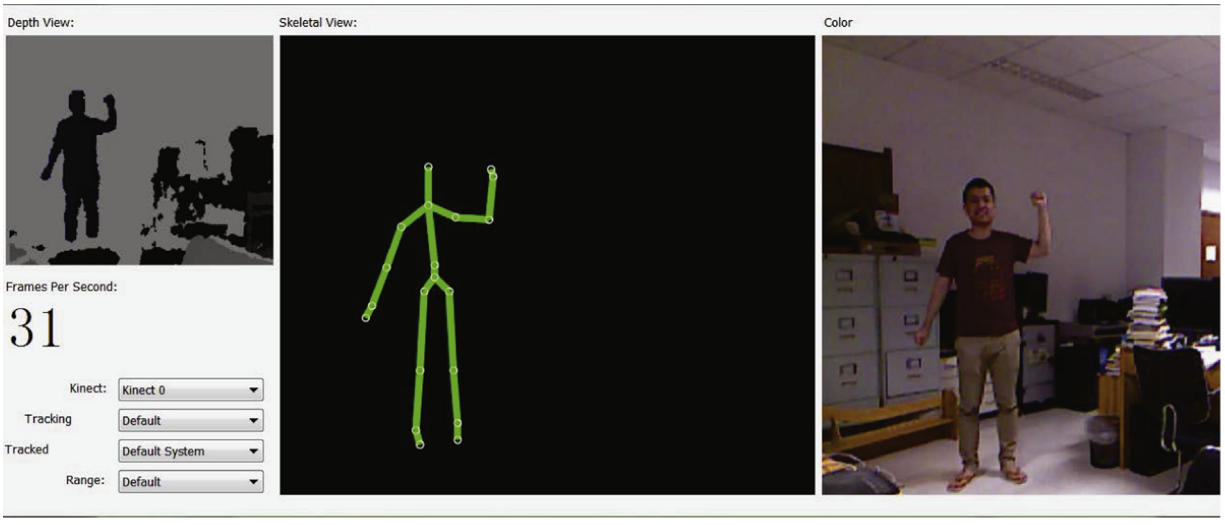

Fig. 3. Depth image, skeleton and color image. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)