Knowledge-Based Systems xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Action recognition based on joint trajectory maps with convolutional neural networks

Pichao Wang[a], Wanqing Li[a], Chuankun Li[b], Yonghong Hou[*,b]

[a] Advanced Multimedia Research Lab, University of Wollongong, Australia
[b] School of Electronic Information Engineering, Tianjin University, China

ARTICLE INFO

Keywords:
Action recognition
Trajectory
Color encoding
Convolutional neural network

ABSTRACT

Convolutional Neural Networks (ConvNets) have recently shown promising performance in many computer vision tasks, especially image-based recognition. How to effectively apply ConvNets to sequence-based data is still an open problem. This paper proposes an effective yet simple method to represent spatio-temporal information carried in 3*D* skeleton sequences into three 2*D* images by encoding the joint trajectories and their dynamics into color distribution in the images, referred to as Joint Trajectory Maps (JTM), and adopts ConvNets to learn the discriminative features for human action recognition. Such an image-based representation enables us to fine-tune existing ConvNets models for the classification of skeleton sequences without training the networks afresh. The three JTMs are generated in three orthogonal planes and provide complimentary information to each other. The final recognition is further improved through multiplicative score fusion of the three JTMs. The proposed method was evaluated on four public benchmark datasets, the large NTU RGB+D Dataset, MSRC-12 Kinect Gesture Dataset (MSRC-12), G3D Dataset and UTD Multimodal Human Action Dataset (UTD-MHAD) and achieved the state-of-the-art results.

1. Introduction

Human action recognition is an important problem in computer vision due to its wide applications in video surveillance, human computer interfaces, robotics, etc. Over the past a few decades, extensive research has been conducted on RGB-based action recognition and the proposed approaches include space-time volume based methods [1–3], space-time trajectory based methods [4–9], appearance-based methods [10,11], motion-encoding based methods [12,13], graph-model based methods [14], key poses based methods [15], kernel-based methods [16] and deep learning based methods [17–19], but accurate recognition of human actions from RGB video sequences is still an unsolved problem. With the advent of easy-to-use and low-cost depth sensors such as MS Kinect sensors, human action recognition from RGB-D (Red, Green, Blue and Depth) data has attracted increasing attention and many applications have been developed [20] in recent years, due to the advantages of depth information over conventional RGB video, e.g. being insensitive to illumination changes and reliable to estimate body silhouette and skeleton [21]. Since the first work [22] reported in 2010, many methods [23–27] have been proposed using specifically hand-crafted feature descriptors extracted from depth. As the extraction of skeletons from depth maps [21] has become increasingly robust, more and more hand-designed skeleton features, such as skeleton joints based methods [28–33], group joints based methods [34–36] and joint dynamics based methods [37–40], have been devised to capture spatial information, and Dynamic Time Warpings (DTWs), Fourier Temporal Pyramid (FTP) or Hidden Markov Models (HMMs) are employed to model temporal information. However, these hand-crafted features are often either shallow, dataset-dependent, or not learned in an end-to-end fashion [41]. Recently, with the development of neural networks and its wide applications [42–44], Recurrent Neural Networks (RNNs) [45–49] have also been adopted for action recognition from skeleton data. RNNs tend to overemphasize the temporal information especially when the training data is not sufficient, leading to overfitting. Up to date, it remains unclear how skeleton sequences could be effectively represented and fed to deep neural networks for recognition. For example, one can conventionally consider a skeleton sequence as a set of individual frames with some form of temporal smoothness, or as a subspace of poses or pose features, or as the output of a neural network encoder. Which one among these and other possibilities would result in the best representation in the context of action recognition is not well understood.

In this paper, we present an effective yet simple method that represents both spatial configuration and dynamics of joint trajectories

* Corresponding author.
E-mail addresses: pw212@uowmail.edu.au (P. Wang), wanqing@uow.edu.au (W. Li), chuankunli@tju.edu.cn (C. Li), houroy@tju.edu.cn (Y. Hou).

https://doi.org/10.1016/j.knosys.2018.05.029
Received 2 October 2017; Received in revised form 19 May 2018; Accepted 21 May 2018

into three texture images through color encoding, referred to as Joint Trajectory Maps (JTMs), as the input of ConvNets for action recognition. Such image-based representation enables us to fine-tune existing ConvNets models trained on ImageNet for classification of skeleton sequences without training the whole deep networks afresh. The three JTMs are complimentary to each other, and the final recognition accuracy is improved largely by a late score fusion method. One of the challenges in action recognition is how to properly model and use the spatio-temporal information. The commonly used bag-of-words model often ignores temporal information. On the other hand, HMMs or RNNs based methods are likely to overstress the temporal information. The proposed method addresses this challenge in a novel way by encoding as much spatio-temporal information as possible (without a need to decide which one is important and how important it is) into images, and employing ConvNets to learn the discriminative one. Consequently, the proposed method outperformed the start-of-the-art methods on popular benchmark datasets.

The main contributions of this paper include:

- A compact, effective yet simple image-based representation is proposed to represent the spatio-temporal information carried in the 3*D* skeleton sequences into three 2*D* images by encoding the dynamics of joint trajectories into three complementary Joint Trajectory Maps.
- To overcome the drawbacks of ConvNets not being rotation-invariant, and to make the proposed method suitable for cross-view action recognition, it is proposed to rotate the skeleton data to not only mimic the multiple views but also to augment data effectively for training.
- The proposed method was evaluated on four popular public benchmark datasets, namely, the large NTU RGB+D Dataset [48], MSRC-12 Kinect Gesture Dataset (MSRC-12) [50], G3D Dataset [51] and UTD Multimodal Human Action Dataset (UTD-MHAD) [52], and achieved the state-of-the-art recognition results.

This paper is an extension of the works presented in [53,54]. Unlike [53,54] where skeletons are assumed to have been sufficiently sampled and discrete joints are drawn onto images using a pen whose size is properly set, this paper employs joint trajectories and proposes to rotate skeletons to mimic multiple views for cross-view action recognition and data augmentation. In addition, this paper adopts multiplicative score fusion to improve the final recognition accuracy. Extensive experiments and detailed analysis are also presented in this paper.

The rest of this paper is organized as follows. An overview of related works is given in Section 2. Details of the proposed method are described in Section 3. Evaluation of the proposed method on four datasets and analysis of the results are reported in Section 4. Section 5 concludes the paper with remarks.

2. Related work

An extensive review on RGB-D based action recognition is beyond the scope of this paper. Readers are referred to [55–57] for a comprehensive survey. In this section, the work related to the proposed method is briefly reviewed, including skeleton-based 3D action representation and deep learning based action recognition.

2.1. Skeleton-based 3D action representation

Skeleton based 3D action representation can be generally divided into three categories [56]: joints, groups of joints , and joint dynamics. Joint representation captures the correlation of the body joints by extracting spatial descriptor [28,58–61], geometric descriptor [32–34,38,40,62] or key poses [29,63–65]. The groups of joints aim to detect the discriminative subsets of joints to differentiate actions. Methods such as [34,63,66–69] focus on mining the subsets of most discriminative joints or consider the correlation of predefined subsets of joints.

Joint dynamics focuses on modeling the dynamics of either subsets or all joints of a skeleton. In [37], 3D trajectories of joints are projected into three 2D trajectories, and histogram of oriented displacement is calculated to describe the three 2D trajectories, with each displacement in the trajectory voting its length in the histogram of orientation angles. Chaudhry et al. [35] divided the fully body skeleton into several body parts represented by joints, including the upper body, lower body, left/right arms and left/right legs. A shape context feature is computed by considering the directions of a set of equidistant points sub-sampled over the segments of each body part. A skeleton sequence is finally represented as a set of time series of features such as position, tangent and shape context feature. These time series are further divided into several temporal scales, and each individual feature series is modeled using a linear dynamic system. The estimated parameters of all series are used to describe the dynamics of the skeleton sequence. In [31], a skeleton sequence is modeled as a continuous and differentiable function of the body joint locations over time. The local 3D body pose is characterized by the current joint locations and differential properties like speed and acceleration of the joints. Slama et al. [70] represented each action sequence as a linear dynamic system that produces 3D joint trajectories. Autoregressive moving average model was adopted to represent the dynamics by means of an observability matrix which embeds the parameters of the model. In [71], the dynamic forest model was proposed and a set of autoregressive trees was adopted. Each node in the probabilistic autoregressive tree stores a multivariate normal distribution with a fixed covariance matrix, and the set of Gaussian posteriors estimated by the forest are used to calculate the forest posterior. Shao and Li [72] proposed to use a class of integral invariants to describe motion trajectories by calculating the line integral of a class of kernel functions at multiple scales along the motion trajectory. In [40], the authors represented the 3D coordinates of joints and their changes over time as a trajectory in the Riemannian manifold, and the action recognition is formulated as the problem of computing the similarity between the shape of trajectories. In this paper, we propose to use color to encode the dynamics of trajectories, and model the spatial-temporal information carried in a skeleton sequence through shape and textures. ConvNets are used to learn deep hierarchy features.

2.2. Deep leaning based action recognition

The exiting deep learning approaches to action recognition can be generally divided into four categories based on how an input sequence is represented and fed to a deep neural network. The first category considers a video either as a set of still images [73] or as a short and smooth transition between similar frames [74,75], and each color channel of the images is fed to one channel of a ConvNet. Although suboptimal, considering the video as a bag of static frames gives reasonable results. The second category is to represent a video as a volume and extends ConvNets to a third, temporal dimension [76,77] replacing 2D filters with 3D ones. So far, this approach has produced little benefits, probably due to the lack of annotated training data. The third category is to treat a video as a sequence of images and feed the sequence to a RNN [45–48,78,79]. An RNN is typically considered as memory cell, which are sensitive to both short as well as long term patterns. It parses the video frames sequentially and encodes the frame-level information in their memory. However, using RNNs has not given an improvement over temporal pooling of convolutional features [73] or over hand-crafted features. The last category is to represent a video in one or multiple compact images and adopt available trained ConvNet architectures for fine-tuning, such as dynamic images based representation for RGB video [19], depth motion maps based representation for depth modality [80,81], or texture images based representation for skeleton sequence [53,54]. This approach has achieved promising results on many RGB and depth/skeleton datasets. The