Computer Vision and Image Understanding 154 (2017) 108–126

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

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

Cross-view human action recognition from depth maps using spectral graph sequences

Tommi Kerola*, Nakamasa Inoue, Koichi Shinoda

Department of Computer Science, Tokyo Institute of Technology, Tokyo 152–8552, Japan

ARTICLE INFO

Article history:
Received 25 March 2016
Revised 18 August 2016
Accepted 16 October 2016
Available online 17 October 2016

Keywords:
Human action recognition
Depth cameras
Spectral graph theory
Graph signal processing
Graph wavelets
Wavelet transform

ABSTRACT

We present a method for view-invariant action recognition from depth cameras based on graph signal processing techniques. Our framework leverages a novel graph representation of an action as a temporal sequence of graphs, onto which we apply a spectral graph wavelet transform for creating our feature descriptor. We evaluate two view-invariant graph types: skeleton-based and keypoint-based. The skeleton-based descriptor captures the spatial pose of the subject, whereas the keypoint-based is able to capture complementary information about human-object interaction and the shape of the point cloud. We investigate the effectiveness of our method by experiments on five publicly available datasets. By the graph structure, our method captures the temporal interaction between depth map interest points and achieves a 19.8% increase in performance compared to state-of-the-art results for cross-view action recognition, and competing results for frontal-view action recognition and human-object interaction. Namely, our method results in 90.8% accuracy on the cross-view N-UCLA Multiview Action3D dataset and 91.4% accuracy on the challenging MSRAction3D dataset in the cross-subject setting. For human-object interaction, our method achieves 72.3% accuracy on the Online RGBD Action dataset. We also achieve 96.0% and 98.8% accuracy on the MSRActionPairs3D and UCF-Kinect datasets, respectively.

1. Introduction

We live in a world where machines are able to either aid or completely replace humans in a large variety of tasks. Most such tasks are quite trivial and monotonic, but thanks to the advent of machine learning, we are at the verge of being able to demand satisfying performance even for more complex tasks. One such task is action recognition. If machines could robustly recognize and interpret human actions and gestures, the benefits would be vast for a number of areas, including games, health care and the security industry.

Classic approaches to action recognition based on simple color images face numerous difficulties due to intra-class variations of actions, background clutter and illumination variations. However, thanks to the emergence of cheap and affordable depth maps with devices such as the Microsoft Kinect, there has been a recent increase in research using 3D features (Han et al., 2013). Leveraging 3D cameras solves the problem of separating the action subject from the video background, and also eliminates irrelevant information such as illumination variance. Recently, due to the work of Shotton et al. (2013), we have access to low-dimensional skeletons mapped to the human body. Out of the box, these skeletons are much more discriminative than the raw high-dimensional RGB-D data and allow the development of efficient methods for action recognition. However, while the 3D skeletons provide means of alleviating the action recognition task, they also provide new challenges due to unstable joint positions resulting from tracking errors in the noisy depth maps.

A recurring question in machine learning is the one of how to best represent objects for handling the pattern learning task. Generally, the approaches to this problem can be divided into two: statistical and structural (Bunke and Riesen, 2012). While statistical methods have received a great deal of attention in the past years, we ask ourselves if objects are not better represented by an explicit structure suitable to the task at hand. Actions are typically defined by a sequence of interactions between several interest points (Johansson, 1973). E.g. “draw circle”:

1. Move hand towards left side of waist.
2. Move hand up.
3. Move hand down towards right side of waist.
4. Move hand down towards feet.

Naturally, a good descriptor for action recognition needs to capture interactions between different parts of the body, all of which

* Corresponding author.
E-mail addresses: kerola@ks.cs.titech.ac.jp (T. Kerola), inoue@ks.cs.titech.ac.jp (N. Inoue), shinoda@cs.titech.ac.jp (K. Shinoda).

http://dx.doi.org/10.1016/j.cviu.2016.10.004
1077-3142/

also vary temporally during the duration of the action. While most existing action recognition methods from depth maps capture such interactions (Luo et al., 2013; Oreifej et al., 2013; Wang et al., 2012; Wang and Wu, 2013; Zhao et al., 2013), most of them are inherently view-dependent. That is, their performance depend on the camera angle from which the action was recorded. Cross-view action recognition is the task of recognizing an action independent of the camera angle used for recording the video. For RGB videos, this has previously been explored to some extent (Farhadi and Tabrizi, 2008; Li et al., 2012; Li and Zickler, 2012; Parameswaran and Chellappa, 2006; Rahmani and Mian, 2015; Rao et al., 2002; Weinland et al., 2007; 2006; Yilmaz and Shah, 2005; Zhang et al., 2013; Zheng and Jiang, 2013). For depth maps, however, the number of methods that apply to cross-view action recognition from pure 3D data are much fewer (Rahmani et al., 2014; Wang et al., 2014a; Xia et al., 2012). This despite the added advantage of being able to perform action recognition without compromising the identity of the user, which is essential for health care applications.

In this work, we consider to use graphs to represent actions due to the following reasons. First, a graph provides a natural structure for representing interactions between interest points. Furthermore, since graphs naturally capture pair-wise information, a graph-based representation is inherently view-invariant provided that this holds for the signal defined on the vertices. This is our motivation for exploring the usage of graphs for action recognition.

In real life problems, graphs can be found everywhere. They occur in forms of *e.g.*social- and transportation networks, finite state machines, and also in domains such as brain fMRI and computer graphics (Shuman et al., 2013). Recent approaches for using graphs in machine learning include graph kernels (Bunke and Riesen, 2011; Hermansson et al., 2013; Shervashidze et al., 2011; Stumm et al., 2016; Zhu et al., 2006), generalizations of signal processing frameworks to the graph domain (Shuman et al., 2013; Sandryhaila and Moura, 2013), and also graph wavelets (Coifman and Maggioni, 2006; Crovella and Kolaczyk, 2003; Hammond et al., 2011; Narang and Ortega, 2012; Ram et al., 2011).

Graph signal processing allows signal propagation that follows the natural structure of objects, and applications include edge-aware image processing (Narang et al., 2012), depth video coding (Kim et al., 2012), image compression (Sandryhaila and Moura, 2012), anomaly detection in wireless sensor networks (Egilmez and Ortega, 2014), bridge structure health monitoring (Chen et al., 2014), brain functional connectivity analysis (Leonardi and Van De Ville, 2011) and mobility pattern prediction (Dong et al., 2013).

Our interest in using graph signal processing for human action recognition lies in the graph frequency information it is able to provide. As our results will show in this paper, using generalizations of wavelet transforms to graphs allows us to capture multi-scale information about the interactions between depth map interest points along with their temporal propagation, leading to an efficient method for classifying a wide range of actions.

In this paper, we propose a system for view-invariant depth map action recognition based on graph signal processing techniques. Our framework leverages a novel graph representation of an action as a temporal sequence of graphs. Specifically, our method takes depth map interest points and embeds these on an augmented graph describing said points' temporal progression. Extending a preliminary study on this subject (Kerola et al., 2014), we investigate two types of interest points:

- Tracked skeleton joints, which capture subject pose and provides a semantic labeling of body parts.
- Spatio-temporal keypoints, which capture human-object interaction and other fine intrinsic detail.

We define view-invariant graph signals based on the above interest points, and we represent them using a novel graph representation that is shown to out-perform more classic representations, such as bag-of-words (BoW) (Salton and McGill, 1986) combined with a support vector machine (SVM) (Chang and Lin, 2011). Particularly, we leverage the spectral graph wavelet transform (SGWT) framework of Hammond et al. (2011) for creating a multi-scale representation of the interest points. Graph wavelets capture information about a signal at different scales, in several dimensions on the augmented temporal graph; both between interest points and along time. Further, spectral graph wavelets offer more flexibility than classical wavelets due to the freedom of graph design. To capture the sequential behavior of actions, we utilize a temporal pyramid pooling scheme (Gowayyed et al., 2013; Luo et al., 2013; Wang et al., 2012) on the wavelet coefficients. This improves over approaches that consider only global information (Li et al., 2010; Yang et al., 2012), since it allows us to capture differently segmented levels of temporal dependencies. Classification is finally performed using an off-the-shelf SVM.

Our proposed method has the following advantages:

- The underlying graph has an explicit block sparsity structure, which we exploit to create a memory-efficient algorithm for calculating the SGWT (see Section 4.3.1).
- The feature's underlying spectral basis is mathematically well defined (Hammond et al., 2011), enabling analysis about each part of the descriptor. On the contrary, methods based on *e.g.*sparse coding (Luo et al., 2013) or deep learning (Rahmani and Mian, 2015) produce bases that are not easily analyzable (see Section 4.9).
- For skeleton-based graphs, the number of interest points N is small, making the method efficiently computable in $\mathcal{O}(TN)$ time, where T is the number of frames, making it more computationally efficient than approaches that rely on solving heavy optimization problems (Luo et al. 2013; Wang and Wu 2013) (see Section 4.7).
- For keypoint-based graphs, the descriptor is shown to capture more information than a baseline BoW-representation, which makes our method perform better using our spectral representation (see Section 5).

While this paper focuses on recognition of actions, the framework can in general be applied to any time series of graphs.

The paper is organized as follows. Section 2 reviews related research in action recognition and graph signal processing. Section 3 discusses how to represent actions as graphs. Our proposed method is then shown in Section 4, with related experiments in Section 5. Section 6 finally concludes the paper.

1.1. Notation

We use lower-case bold letters $\mathbf{a} = [a(1), \ldots, a(n)]^T$ to denote vectors, and $a(i)$ denotes the ith element of a vector. We use upper-case bold letters **A, B, C** to denote matrices, with A(i, j) referring to the element at the ith row and jth column of **A**. Let $\mathbf{a}_n$ denote the nth vector in a set of vectors. We use $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{W})$ to denote an undirected graph with vertex set $\mathcal{V} = \{v_i\}$ and edge set $\mathcal{E} = \{e_k : e_k = (v_i, v_j) \Leftrightarrow v_i \sim v_j;\ v_i, v_j \in \mathcal{V}\}$ and $v_i \sim v_j$ denotes that vertices i, j are connected by an edge. The weight matrix **W** stores the weight of an edge (v_i, v_j) in entry W(i, j).

2. Related work

2.1. 3D action recognition

The advent of cheap 3D cameras such as the Kinect has enabled a great performance increase for action recognition tasks (Li et al. 2010). The availability of RGB-D data has considerably eased the task of segmenting an actor from its background; something that