Graph Based Skeleton Modeling for Human Activity Analysis

Kao, J.-Y.; Ortega, A.; Tian, D.; Mansour, H.; Vetro, A.

TR2019-037 June 28, 2019

Abstract

Understanding human activity based on sensor information is required in many applications and has been an active research area. With the advancement of depth sensors and tracking algorithms, systems for human motion activity analysis can be built by combining off-the-shelf motion tracking systems with application-dependent learning tools to extract higher semantic level information. Many of these motion tracking systems provide raw motion data registered to the skeletal joints in the human body. In this paper, we propose novel representations for human motion data using the skeletonbased graph structure along with techniques in graph signal processing. Methods for graph construction and their corresponding basis functions are discussed. The proposed representations can achieve comparable classification performance in action recognition tasks while additionally being more robust to noise and missing data.

IEEE International Conference on Image Processing (ICIP)

GRAPH BASED SKELETON MODELING FOR HUMAN ACTIVITY ANALYSIS

Jiun-Yu Kao[1] *Antonio Ortega*[1] *Dong Tian*[3] *Hassan Mansour*[2] *Anthony Vetro*[2]

[1] Department of Electrical and Computer Engineering, University of Southern California,
3740 McClintock Ave., Los Angeles, CA 90089, USA
[2] Mitsubishi Electric Research Labs (MERL),
201 Broadway, Cambridge, MA 02139, USA
[3] InterDigital,
4 Research Way, Suite 300, Princeton, NJ 08540, USA

ABSTRACT

Understanding human activity based on sensor information is required in many applications and has been an active research area. With the advancement of depth sensors and tracking algorithms, systems for human motion activity analysis can be built by combining off-the-shelf motion tracking systems with application-dependent learning tools to extract higher semantic level information. Many of these motion tracking systems provide raw motion data registered to the skeletal joints in the human body. In this paper, we propose novel representations for human motion data using the skeleton-based graph structure along with techniques in graph signal processing. Methods for graph construction and their corresponding basis functions are discussed. The proposed representations can achieve comparable classification performance in action recognition tasks while additionally being more robust to noise and missing data.

Index Terms— Human activity analysis, graph-based representation, motion capture data, 3D action recognition

1. INTRODUCTION

Understanding human motion continues to be a challenging and active area of research. Recently, cost-effective depth sensors, such as Kinect, combined with powerful real-time tracking algorithms [1, 2], provide fairly accurate 2D or 3D positions of skeletal joints, allowing for human activity analysis systems to infer or predict specific actions. In this work, we focus on using the skeleton-based motion data, i.e., 2D or 3D coordinates associated with each skeletal joint, as the captured motion data and aim to represent and analyze this information for improved activity recognition.

Assuming that skeleton-based motion data is available, recognition systems typically extract descriptive and compact information, i.e., representations or features, to characterize the attributes in human motion. State-of-the-art approaches to extract representations are mostly data-driven. For example, in principal component analysis (PCA) methods [3, 4], the representation, i.e., the representations are learned from the raw motion data. PCA-based representations are efficient in terms of energy compaction, but the principal components need to be recomputed for new datasets, thus requiring complex retraining to generalize to new datasets. Moreover, they are sensitive to noise and missing data, which is common in skeleton-based motion tracking systems. PCA-based approaches do not explicitly consider the spatial dependency among body joints. Since actions are performed by human bodies, and physical restrictions on their motion are known *a priori*, incorporating knowledge about the skeletal structure, rather than relying solely on data, can be beneficial, especially in terms of robustness to noise and missing data. The main challenge in developing such representations is the irregularity in the skeletal structure and its corresponding motion, which can potentially be tackled by leveraging graph structure derived from the skeleton along with graph signal processing approaches, where notions of frequency derived from spectral graph theory are used to process data in irregular domains [5, 6, 7].

Approaches for human motion analysis using graphs based on natural skeletal structure have been proposed in the past. For example, in [8, 9], an undirected skeletal graph is constructed and motion data are regarded as signals on such graph. Furthermore, the development of graph convolutional networks (GCNs) [10, 11] has made it possible to use graph-based data directly for classification. Although the GCN-based approaches [12, 13, 14] utilize graphs to model prior knowledge about human skeleton, methods to extract representations are still data-driven, e.g., learning the best graph filtering functions [12, 13, 14] or learning the best graph structure [13, 14] from data. These data-driven approaches may have advantages in terms of discriminating between different action categories, and show superior performance in the context of action recognition, but they cannot be easily generalized across datasets, other than by fully retraining the system for the new tasks.

As an alternative to data-driven techniques, we instead focus on model-based approaches to *construct* the representations. We propose graph-based motion representations that start with a skeletal-temporal graph and then apply an existing transform, such as graph Fourier transform (GFT), to the graph signal defined on the constructed graph. A key benefit of this construction is that it allows us to interpret the actions using the spectrum and basis vectors of the constructed graph. Compared to PCA-based methods where transformation is learned from data, the transformation we construct does not depend on data but on knowledge about skeleton, leading to better interpretation, robustness to noisy and missing data, and easier generalization across datasets and tasks. Moreover, unlike GCN-based methods where both the graph and graph filters are learned from data, our proposed approach utilizes a fixed skeletal-temporal graph with a known graph transformation, so that the graph itself is not dependent on data. Our analysis on basis functions of the skeletal graph and the resulting interpretations can also provide insights on why GCN works. Furthermore, we demonstrate via action recognition experiments that the proposed graph-based representation achieves better robustness to noise and missing data compared to data-driven approaches such as PCA-based methods. It is worth noting that, in order to evaluate robustness to noise of representations, we investigate the characteristics of noise in skeleton-based motion data and propose a joint-dependent noise model for generating artificial noise to be added to skeleton data.

The remainder of this paper is organized as follows. In Section 2, we present the proposed framework to construct graph-based representations for motion data, together with the interpretations it provides. We further validate that the proposed representations are more robust to noise and missing data in the context of 3D action recognition in Section 3. Section 4 concludes this paper.

2. PROPOSED GRAPH-BASED REPRESENTATIONS

Suppose that 3D skeleton-based motion data are available with N_s nodes representing tracked body joints (or some predefined keypoints). For each motion sequence, the i-th tracked joint is associated with its estimated 3D position at frame t, denoted as $\mathbf{p}_{t,i} = \begin{pmatrix} x_i^{(t)} & y_i^{(t)} & z_i^{(t)} \end{pmatrix}$, where $i \in \{1, \cdots, N_s\}$ and $t \in \{1, \cdots, T\}$.

Graph construction: We first model the human skeletal structure as a fixed undirected *skeletal graph* $\mathcal{G}_s = \{\mathcal{V}_s, \mathcal{E}_s, \mathbf{W}\}$ with the vertex set $\mathcal{V}_s = \{v_1, v_2, \cdots, v_{N_s}\}$ corresponding to the N_s tracked body joints. The edge set $\mathcal{E}_s$ consists of undirected edges with unity weights, which are specified in $\mathbf{W}$. $\mathcal{E}_s$ is decided based on knowledge about human skeleton as follows: v_i is connected to v_j with a unity weight only if there exists a physical limb directly connecting the i-th and j-th body joint. In this way, the constructed skeletal graph $\mathcal{G}_s$ captures the physical connectivity between body parts.

Once the graph is defined, the combinatorial graph Laplacian matrix $\mathbf{L}$ is defined as $\mathbf{L} \equiv \mathbf{D} - \mathbf{W}$ while the normalized graph Laplacian matrix $\mathcal{L}$ is defined as

$$\mathcal{L} \equiv \mathbf{D}^{-\frac{1}{2}}\mathbf{L}\mathbf{D}^{-\frac{1}{2}} = \mathbf{I} - \mathbf{D}^{-\frac{1}{2}}\mathbf{W}\mathbf{D}^{-\frac{1}{2}} \tag{1}$$

where $\mathbf{I}$ is the identity matrix and $\mathbf{D}$ is the degree matrix, i.e., $\mathbf{D}_{ii} = \sum\limits_{j \neq i} \mathbf{W}_{ij}$. Since $\mathcal{L}$ is real and symmetric when $\mathcal{G}_s$ is undirected, its eigendecomposition can be shown to be:

$$\mathcal{L} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^{\mathsf{T}} = \sum_{i=1}^{N} \lambda_i \mathbf{u}_i \mathbf{u}_i^{\mathsf{T}} \tag{2}$$

where λ_i is the i-th smallest eigenvalues of $\mathcal{L}$ corresponding to eigenvector $\mathbf{u}_i$, $\mathbf{\Lambda} \equiv \operatorname{diag}(\lambda_i)$, $\mathbf{U} \equiv \begin{pmatrix} \mathbf{u}_1 & \mathbf{u}_2 & \cdots & \mathbf{u}_N \end{pmatrix}$ and $\mathbf{U}\mathbf{U}^{\mathsf{T}} = \mathbf{I}$. That is, the eigenvectors form an orthonormal basis and the set of eigenvalues $\sigma(\mathcal{G}_s) \equiv \{\lambda_1, \cdots, \lambda_N\}$ is the spectrum of graph $\mathcal{G}_s$. The eigenvectors of $\mathcal{L}$ associated with larger eigenvalues correspond to elementary graph signals exhibiting greater variation across connected vertices. Thus $\mathbf{U}$ can be viewed as providing a frequency decomposition analogous to the classical Fourier transform for 1D signals (see [5, 6, 7] for more details). The matrix of eigenvectors $\mathbf{U}$ defines the Graph Fourier Transform (GFT). For any graph signal $\mathbf{x} \in \mathbb{R}^N$, the GFT is defined as

$$\tilde{\mathbf{x}} = \mathbf{U}^{\mathsf{T}}\mathbf{x}. \tag{3}$$

Fig. 1 illustrates some of the GFT basis vectors of $\mathcal{G}_s$, which we can be interpreted in terms of the corresponding elementary motion. For example, the 4$^{\text{th}}$ vector can correspond to typical motion while walking since bilateral symmetry is a well-known characterization in normal human gait.

We can further construct the *skeletal-temporal graph* $\mathcal{G}_{st}$ with N_t temporal nodes by taking the graph Cartesian product of a skeletal graph $\mathcal{G}_s$ and an unweighted temporal line graph $\mathcal{G}_t$ with N_t vertices, i.e., $\mathcal{G}_{st} = \mathcal{G}_s \square \mathcal{G}_t$. Fig. 2 shows an example of constructing a skeletal-temporal graph with 2 temporal nodes and 15 joints.

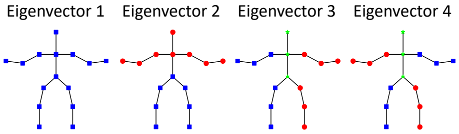

Fig. 1. First four GFT basis vectors $\mathbf{u}_1, \cdots, \mathbf{u}_4$ of a 15-node skeletal graph. Blue square: positive value. Red dot: negative value. Green pentagram: zero.

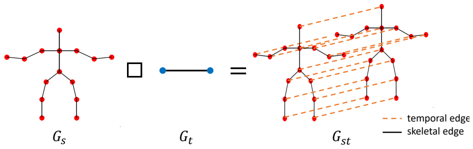

Fig. 2. Example of constructing a skeletal-temporal graph with $N_s = 15$, $N_t = 2$.

The GFT basis of a skeletal-temporal graph can be interpreted by noting that they are closely related to the GFT bases of $\mathcal{G}_s$ and $\mathcal{G}_t$. More formally, assume that $\mathbf{x}$ is one GFT basis vector of $\mathcal{G}_s$ ($\mathbf{x}$ as an eigenvector of $\mathbf{W}_{\mathcal{G}_s}$ with eigenvalue λ) and $\mathbf{y}$ is one GFT basis vector of $\mathcal{G}_t$ ($\mathbf{y}$ as an eigenvector of $\mathbf{W}_{\mathcal{G}_t}$ with eigenvalue μ), then since

$$\begin{aligned} &\mathbf{W}_{\mathcal{G}_{st}} \cdot (\mathbf{x} \otimes \mathbf{y}) \\ &= \left(\mathbf{W}_{\mathcal{G}_s} \otimes \mathbf{I}_{N_t}\right)(\mathbf{x} \otimes \mathbf{y}) + \left(\mathbf{I}_{N_s} \otimes \mathbf{W}_{\mathcal{G}_t}\right)(\mathbf{x} \otimes \mathbf{y}) \\ &= \mathbf{W}_{\mathcal{G}_s}\mathbf{x} \otimes \mathbf{I}_{N_t}\mathbf{y} + \mathbf{I}_{N_s}x \otimes \mathbf{W}_{\mathcal{G}_t}\mathbf{y} \\ &= (\lambda + \mu)(\mathbf{x} \otimes \mathbf{y}) \end{aligned} \tag{4}$$

where $\otimes$ denotes the Kronecker product, we can see that $\mathbf{x} \otimes \mathbf{y}$ is one GFT basis vector of $\mathcal{G}_{st}$. That is, the GFT basis vectors of the skeletal-temporal graph can be derived as the Kronecker product between basis vectors of typical skeletal graph and basis vectors of temporal line graph. Fig. 3 provides an illustrative example.

Graph signals: Any spatial-temporal cube of length N_t in the motion sequence can be regarded as a graph signal residing on the skeletal-temporal graph. Specifically, a graph signal $\mathbf{f}_d^{(t)} \in \mathbb{R}^{N_s \times N_t}$ can be defined on this skeletal-temporal graph $\mathcal{G}_{st}$ when having $\mathbf{f}_d^{(t)}(i, s) = d_i^{(t+s-1)}$, where $i \in \{1, \cdots, N_s\}$, $s \in \{1, \cdots, N_t\}$, for any $d \in \{x, y, z\}$ and $t \in \{1, \cdots, T - N_t + 1\}$.

Proposed representations: Once the skeletal-temporal graph is constructed, we can use transforms associated with the graph to represent motion data. Specifically, for *GFT-based representation*, we apply the GFT to graph signals as $\widetilde{\mathbf{f}_d^{(t)}} = \mathbf{U}^{\mathsf{T}}\mathbf{f}_d^{(t)}$ as in (3) and utilize the transform coefficients $\widetilde{\mathbf{f}_d^{(t)}}$ as the representation.

3. EXPERIMENTS: 3D ACTION RECOGNITION

3.1. Feature Design

As the proposed representations are constructed frame-wise, given a motion sequence and its associated frame-wise representation, we need to choose a temporal model to capture the temporal dynamics. In our experiments, we adopt temporal pyramid matching (TPM) [15] to model the dynamics in the sequence of frame-wise representations, but alternative temporal models could be used

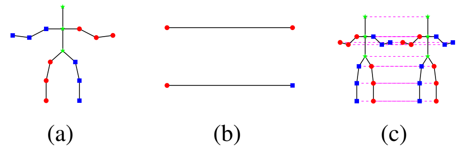

Fig. 3. (a) A GFT basis vector of $\mathcal{G}_s$. (b) Two GFT basis vectors of $\mathcal{G}_t$ with $N_t = 2$. (c)(d) Two GFT basis vectors of $\mathcal{G}_{st}$, each of which is Kronecker product between (a) and one of (b).

as well. Specifically, given a motion sequence with T frames, we extract the frame-wise GFT-based representation $\widetilde{\mathbf{f}_d^{(t)}}$. Assume GFT-based representations with a skeletal graph are used, a coefficient matrix $\mathbf{C} \in \mathbb{R}^{T\times 3D}$ can be constructed where the i-th row of $\mathbf{C}$ is $\left(\widetilde{\mathbf{f}_x^{(i)}}^{\mathsf{T}}, \widetilde{\mathbf{f}_y^{(i)}}^{\mathsf{T}}, \widetilde{\mathbf{f}_z^{(i)}}^{\mathsf{T}}\right)$ with D as the dimension of $\widetilde{\mathbf{f}_d^{(i)}}$. Then a pooling function is defined to apply column-wise pooling for a sub-block of $\mathbf{C}$. Here we adopt a *mean* pooling function $p : \mathbb{R}^{r\times 3D} \to \mathbb{R}^{1\times 3D}$, which takes column-wise mean for a block of coefficients. Furthermore, we apply this pooling function to sub-blocks of $\mathbf{C}$ of different sizes, which can capture the temporal order of actions spanning different duration. The maximum pyramid level needs to be specified, denoted as M. For the pyramid level $m \leq M$, we first uniformly divide $\mathbf{C}$ into a set of non-overlapping sub-blocks $\{\mathbf{B}_i\}$ so that $\mathbf{C} = \left(\mathbf{B}_1^{\mathsf{T}}, \cdots, \mathbf{B}_{2^{m-1}}^{\mathsf{T}}\right)^{\mathsf{T}}$. The feature vector for this pyramid level is then computed as $\mathbf{z}_m = \left(p\left(\mathbf{B}_1\right), \cdots, p\left(\mathbf{B}_{2^{m-1}}\right)\right)$. Finally, the feature vector for this motion sequence is given by the concatenation of the feature vectors of all the pyramid levels, i.e., $\left(\mathbf{z}_1, \cdots, \mathbf{z}_M\right)$. This temporal pooling scheme (TPM) will be used in the following experiments to extract the feature vector for each motion sequence.

3.2. Datasets & Experimental Settings

We evaluate the proposed representations in the context of action recognition using two public datasets: **MSR-Action3D [16]** and **UTKinect-Action3D [17]**. Both datasets were captured by a depth sensor, e.g., Kinect, and the 3D positions of 20 skeletal joints are provided. MSR-Action3D contains 20 action categories which results in 557 motion sequences while UTKinect-Action3D contains 10 actions which leads to 199 motion sequences in total.

For both datasets, we adopt the cross-subject evaluation scheme, where the motion sequences from half of the subjects are used for training while the other half are used for testing. The unweighted skeletal-temporal graph is constructed for each dataset based on the number of tracked skeletal joints, i.e., $N_s = 20$ for both datasets. The number of temporal nodes in the skeletal-temporal graph is selected by cross validation. Once the graph is constructed, we compute the proposed GFT-based representations for frames in each sequence and the TPM scheme mentioned in Section 3.1 is adopted to generate the final feature vector for each sequence. For TPM, the maximum pyramid level M is set to 3 in all the experiments.

3.3. Robustness to Noisy Data

In our experiments we add noise at various peak signal-to-noise ratio (PSNR) levels to both datasets and compare the classification accuracy of our proposed approach with that achieved by a PCA-based method. A naive approach to incorporate noise into the simulations would simply consist of selecting some existing model (say, additive white Gaussian noise) and adding noise with equal variance to all motion measurements. Instead, we propose a more realistic model where the noise is joint dependent.

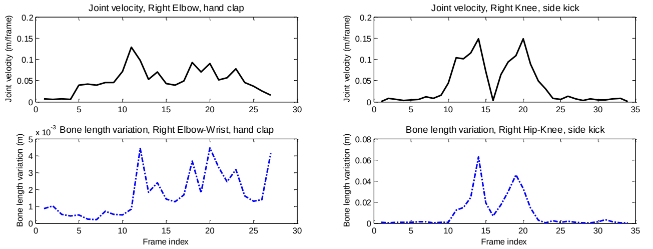

Fig. 4. Examples demonstrate that the level of noise at joint, measured by the variation in the length of attaching bone between consecutive frames, depends on the joint velocity.

We start by analyzing how measurement noise may depend on the specific joint. Because the bone length between a physically connected pair of joints is constant, we first measure the standard deviation of the bone length measurements obtained from the sequence, i.e.,

$$\sigma_{b_k} = \operatorname{std}\left(\left\|\mathbf{p}_{t,i} - \mathbf{p}_{t,j}\right\|_2\right), \tag{5}$$

where b_k is the k-th bone connecting joint i and joint j. Then the standard deviation of noise at each joint is computed as the average over the standard deviations of the lengths of all bones connected to that joint. As for the signal energy, based on the observation that the signal energy should be zero in a static motion, we calculate the peak signal energy E_s by considering the maximum squared norms of all motion vectors at all joints along time, i.e.,

$$E_s = \max_{t\in\{1,\cdots,T-1\},\, i\in\{1,\cdots,N\}} \left\|\mathbf{v}_{t,i}\right\|_2^2, \tag{6}$$

where $\mathbf{v}_{t,i} = \mathbf{p}_{t+1,i} - \mathbf{p}_{t,i}$. Finally, the empirical PSNR can be calculated by taking the ratio between the peak signal energy across all joints and the average noise energy at each joint.

To simulate realistic added noise, we consider a noise model where noise level at each joint is proportional to the moving velocity of that joint. This model is reasonable, as shown in Fig. 4. A joint-dependent model was also proposed in [18] using a different methodology (with manual annotations to obtain the ground truth position).

In the following experiment to validate robustness to noisy data, for each joint i at time t, an independent Gaussian noise with standard deviation as

$$\sigma_{i,t} = \sigma_{psnr} \times \sqrt{\frac{\left\|\mathbf{v}_{t,i}\right\|_2^2}{E_S}} \tag{7}$$

is added to the original data, leading to a noise-corrupted dataset, where σ_{psnr} is decided by the targeted PSNR value. The range of PSNR values for the experiment is selected based on the empirical PSNR in each dataset, which is 59.21dB in MSR-Action3D and 34.17dB in UTKinect-Action3D, averaged across joints. Our proposed GFT representation and PCA are applied to the corrupted data to generate frame-based representations. The same temporal pooling scheme, i.e., temporal pyramid pooling with three pyramid levels, is used for both representations to produce the feature vector for each sequence. Finally, a linear SVM classifier is applied to the feature vectors and the classification accuracy is reported and plotted, as in

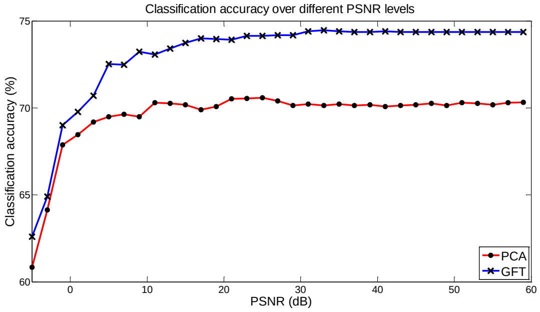

Fig. 5. Classification accuracy over PSNR on MSR-Action3D dataset.

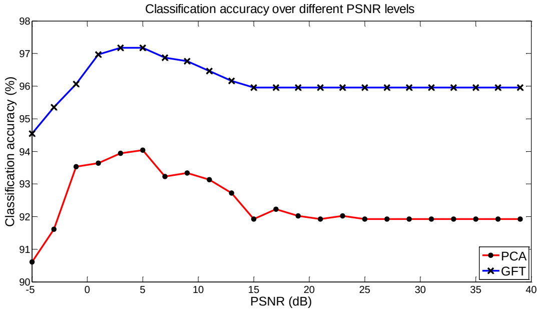

Fig. 6. Classification accuracy over PSNR on the UTKinect-Action3D dataset.

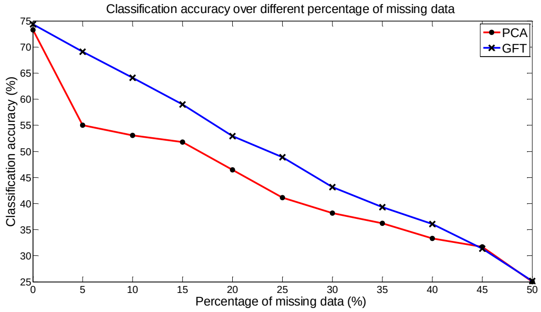

Fig. 7. Classification accuracy over percentage of missing data on MSR-Action3D dataset.

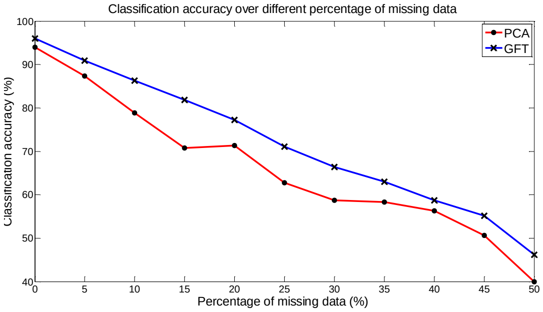

Fig. 8. Classification accuracy over percentage of missing data on the UTKinect-Action3D dataset.

Fig. 5 and 6. The experiment is repeated 10 times for each PSNR level in order to average over the random noise realizations.

Based on the experimental results, we can observe that GFT-based method consistently outperforms the PCA-based method on both datasets when the PSNR is greater than 0, which demonstrates the robustness to noisy data of the proposed graph-based representations.

3.4. Robustness to Missing Data

Next, we evaluate the performance of our proposed schemes in cases where there are missing data (e.g., the position of a joint cannot be obtained at some point in time.) Given a skeleton-based motion dataset, we first synthesize the corresponding corrupted dataset with missing entries. Each entry in the dataset is kept with probability p; otherwise, that entry is thrown away. This error is introduced independently at each joint. Here we only consider the scenario where the percentage of missing data is less than 50%. Based on this corrupted dataset, classification is performed with either PCA-based method or GFT-based method. For PCA-based method, following the conventional approach, we use alternating least squares (ALS) algorithm to jointly learn the principal components and estimated coefficients [19, 20].

For GFT-based method, we propagate the signals based on the predefined skeletal graph. Given each graph signal $\mathbf{f} \in \mathbb{R}^n$, i.e., the motion data per dimension per frame, and the predefined skeletal graph $\mathcal{G}$ with normalized Laplacian as $\mathcal{L}$, $\mathbf{f}$ can be written as $\begin{pmatrix}\mathbf{f}_l & \mathbf{f}_u\end{pmatrix}^\top$ where $\mathbf{f}_l$ represents the observed data while $\mathbf{f}_u$ represents the missing values. $\mathcal{L}$ can then be split accordingly as $\begin{pmatrix}\mathcal{L}_{ll} & \mathcal{L}_{lu} \\ \mathcal{L}_{ul} & \mathcal{L}_{uu}\end{pmatrix}$. We then solve for the optimal $\mathbf{f}_u$ such that $\mathbf{f}^\top\mathcal{L}\mathbf{f}$ is minimized. The closed-form solution is $\hat{\mathbf{f}_u} = -\mathcal{L}_{uu}^{-1}\mathcal{L}_{ul}\mathbf{f}_l$. The propagated signal $\hat{\mathbf{f}} = \begin{pmatrix}\mathbf{f}_l & \hat{\mathbf{f}_u}\end{pmatrix}^\top$ is regarded as the reconstructed data matrix to extract the representation with the proposed GFT-based method described in Section 2.

For each p value, the experiment is repeated 10 times and the averaged classification accuracy is reported for each dataset, see Figs. 7 and 8. We can see that GFT-based method with signal propagation on graph consistently outperforms the PCA-based method with ALS algorithm on both datasets.

4. CONCLUSION

This paper presents a novel framework to construct representations for human motion data by leveraging graph structures. The human skeleton is modeled with a skeletal-temporal graph, where the tracked body joints are the graph vertices and the motion data is the graph signal residing on this graph. Existing graph transforms such as GFT are utilized to extract representations for the captured human motion data. Evaluation of our proposed representations in the context of 3D action recognition demonstrate comparable classification performance compared to conventional PCA-based methods, while providing greater robustness to noisy and missing data. Although not described in detail due to space constraints, it should be mentioned that there is no need to re-compute the transform given a new dataset with the proposed method. As such, the proposed method has significantly lower time complexity to compute the feature representation compared to PCA-based methods, making the proposed method more readily applicable to new datasets.

5. REFERENCES

[1] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake, "Real-time human pose recognition in parts from single depth images," in *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2011, pp. 1297–1304.

[2] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *CVPR*, 2017.

[3] M. Raptis, D. Kirovski, and H. Hoppe, "Real-time classification of dance gestures from skeleton animation," in *Proceedings of the 2011 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, 2011, pp. 147–156.

[4] X. Yang and Y. Tian, "Eigenjoints-based action recognition using nave-bayes-nearest-neighbor," in *CVPR 2012 HAU3D Workshop*. 2012, pp. 14–19, IEEE.

[5] D. I Shuman, S. K. Narang, P. Frossard, A. Ortega, and P. Vandergheynst, "The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains," vol. 30, no. 3, pp. 83–98, 2013.

[6] A. Sandryhaila and J. M. F. Moura, "Big data analysis with signal processing on graphs: Representation and processing of massive data sets with irregular structure," vol. 31, no. 5, pp. 80–90, 2014.

[7] A. Ortega, P. Frossard, J. Kovačević, J. Moura, and P. Vandergheynst, "Graph signal processing: Overview, challenges, and applications," *Proceedings of the IEEE*, vol. 106, no. 5, pp. 808–828, 2018.

[8] J.-Y. Kao, A. Ortega, and S.S. Narayanan, "Graph-based approach for motion capture data representation and analysis," in *Image Processing (ICIP), 2014 IEEE International Conference on*, Oct 2014, pp. 2061–2065.

[9] J.-Y. Kao, M. Nguyen, L. Nocera, C. Shahabi, A. Ortega, C. Winstein, I. Sorkhoh, Y.-c. Chung, Y.-a. Chen, and H. Bacon, "Validation of automated mobility assessment using a single 3d sensor," *Computer Vision - ECCV 2016 Workshops. ECCV 2016. Lecture Notes in Computer Science*, vol. 9914, pp. 162–177, 2016.

[10] J. Bruna, W. Zaremba, A. Szlam, and Y. Lecun, "Spectral networks and locally connected networks on graphs," in *International Conference on Learning Representations (ICLR2014), CBLS, April 2014*, 2014.

[11] M. Defferrard, X. Bresson, and P. Vandergheynst, "Convolutional neural networks on graphs with fast localized spectral filtering," in *Proceedings of the 30th International Conference on Neural Information Processing Systems*, 2016, NIPS'16, pp. 3844–3852.

[12] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *AAAI*, 2018.

[13] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Adaptive Spectral Graph Convolutional Networks for Skeleton-Based Action Recognition," *ArXiv e-prints*, May 2018.

[14] C. Li, Z. Cui, W. Zheng, C. Xu, R. Ji, and J. Yang, "Action-attending graphic neural network," *IEEE Transactions on Image Processing*, vol. 27, pp. 3657–3670, 2018.

[15] P. Wang, Y. Cao, C. Shen, L. Liu, and H. T. Shen, "Temporal pyramid pooling-based convolutional neural network for action recognition," *IEEE Trans. Cir. and Sys. for Video Technol.*, vol. 27, no. 12, pp. 2613–2622, Dec. 2017.

[16] W. Li, Z. Zhang, and Z. Liu, "Action recognition based on a bag of 3d points," in *CVPRW*, 2010.

[17] L. Xia, C. C. Chen, and J. K. Aggarwal, "View invariant human action recognition using histograms of 3d joints," in *CVPRW*, 2012.

[18] S. Obdrzálek, G. Kurillo, F. Ofli, R. Bajcsy, E. Y. W. Seto, H. B. Jimison, and M. Pavel, "Accuracy and robustness of kinect pose estimation in the context of coaching of elderly population," *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 1188–1193, 2012.

[19] F. W. Young, Y. Takane, and J. de Leeuw, "The principal components of mixed measurement level multivariate data: An alternating least squares method with optimal scaling features," *Psychometrika*, vol. 43, no. 2, pp. 279–281, Jun 1978.

[20] T. Hastie, R. Mazumder, J. D. Lee, and R. Zadeh, "Matrix completion and low-rank svd via fast alternating least squares," *J. Mach. Learn. Res.*, vol. 16, no. 1, pp. 3367–3402, Jan. 2015.