MITSUBISHI ELECTRIC RESEARCH LABORATORIES
https://www.merl.com

Application-agnostic spatio-temporal hand graph representations for stable activity understanding

Das, Pratyusha; Ortega, Antonio; Chen, Siheng; Mansour, Hassan; Vetro, Anthony

TR2021-112 October 15, 2021

Abstract

Understanding complex hand actions, such as assembly tasks or kitchen activities, from hand skeleton data is an important yet challenging task. This paper introduces a hand graph-based spatio-temporal feature extraction method which uniquely represents complex hand action in an unsupervised manner. To evaluate the efficacy of the proposed representation, we consider action segmentation and recognition tasks. The segmentation problem involves an assembling task in an industrial setting, while the recognition problem deals with kitchen and office activities. Additionally, for both segmentation and recognition models, we propose notions of stability, which are used to demonstrate the robustness of our proposed approach. We introduce validation loss stability (ValS) and estimation stability with cross-validation (EtS) to analyze robustness of any supervised classification model. The proposed method shows comparable classification performance with state of the art methods, but it achieves significantly better accuracy and stability in a cross-person setting.

IEEE International Conference on Image Processing (ICIP) 2021

Mitsubishi Electric Research Laboratories, Inc.
201 Broadway, Cambridge, Massachusetts 02139

APPLICATION-AGNOSTIC SPATIO-TEMPORAL HAND GRAPH REPRESENTATIONS FOR STABLE ACTIVITY UNDERSTANDING

Pratyusha Das[1], *Antonio Ortega*[1], *Siheng Chen*[2], *Hassan Mansour*[3], *Anthony Vetro*[3]

[1]University of Southern California, Los Angeles, CA, USA
[2]Shanghai Jiao Tong University, Shanghai, China
[3]Mitsubishi Electric Research Labs, Cambridge, MA, USA

ABSTRACT

Using hand skeleton data to understand complex hand actions, such as assembly tasks or kitchen activities, is an important yet challenging task. This paper introduces an unsupervised hand graph-based spatio-temporal feature extraction method. To evaluate the efficacy of the proposed representation, we consider action segmentation and recognition tasks. The segmentation problem involves an assembling task in an industrial setting, while the recognition problem deals with kitchen and office activities. For both tasks, we propose novel notions of stability, loss function stability (LFS) and estimation stability with cross-validation (ESCV), that are used to quantify the robustness of achieved solutions. Our proposed feature extraction leads to classification performance comparable to state of the art methods, while achieving significantly better accuracy and stability in a cross-person setting. The proposed method also outperforms the existing methods in the segmentation task in terms of accuracy and shows robustness to any change in the input hyper-parameters.

Index Terms— Complex activity understanding, Spatio-temporal hand graph, Unsupervised feature extraction, Graph signal processing, Stability analysis.

1. INTRODUCTION

Understanding human activity plays a key role in many areas, such as security, worker monitoring, and human-robot interaction. Developing efficient unsupervised representations for data used in human activity analysis remains an interesting yet difficult problem because of the complex nature of human actions. Human activities such as assembly tasks [1] and food preparation [2] consist of a pre-defined sequence of action units. Analysis of complex activities performed using only hands is challenging due to the similarity of motions in different action units, which makes them hard to distinguish, and localization of the motion to small areas of the body. Complex activity analysis from a video is further complicated when tracking people in a cluttered background [3].

The understanding of complex activity [4] requires a multi-level analysis, starting with the extraction of low level body position information. Recently, low level extraction of detailed 2D/3D positions of the body joints, hands, and face has improved significantly with the emergence of more accurate RGB-D capture devices, e.g., Kinect, and software, e.g., OpenPose [5], which can extract information with low-latency and acceptable accuracy. Given this low level information, a complete task can be divided into smaller sub-tasks depending on the nature of the hand motion and the sub-tasks are then classified according to their semantic meaning. In this paper, we focus on the problem of representation and processing of low level position data so that it can be used for efficient activity segmentation and recognition.

Multiple approaches for action understanding using full-body skeleton data have been proposed in the last decade, including co-occurrence feature learning [6], spatial-temporal graph convolutional networks [7], and spectral graph skeletons [8]. Graph-based approaches for human motion analysis [7, 9, 10] using skeleton data have also gained popularity due to their simplicity and efficiency. In particular, they can provide motion features without prior knowledge of the task. While tasks involving whole body motion have been studied, complex activity understanding using only hand skeleton data has not been thoroughly explored yet. In [7, 11], a neural network-based supervised approach was developed to model dynamic hand skeleton motion and analyze complex activities. However, to achieve good performance this model requires a significant amount of training data, which may not be available in some scenarios. Moreover, such systems may be difficult to use in practice, as retraining and adaptation could be costly (e.g., if the system is moved from one location to another).

In this work, we systematically study hand graphs [12] and complex hand motion in an unsupervised manner. We propose a Graph-based Application-agnostic Feature eXtraction (GrAFX) method for complex action understanding using hand Motion capture (Mocap) data. This feature extraction model is application agnostic and unsupervised. Thus, our method can be used in any hand skeleton based application, from hand action understanding to hand gesture recognition [13]. Our proposed graph-based method extracts temporal and spatial information present in the hand activity sequence. We are particularly interested in developing stable models [14], which are always preferable since they maintain consistency in their performance in varying datasets and conditions. While techniques to analyze stability have been developed [15, 16], none of them provides a metric that can be used to estimate stability for arbitrary machine learning models, including state of the art methods based on neural networks. Additionally, while a estimation stability with cross-validation (ESCV) has been proposed for problems such as a Lasso based optimization [17], similar techniques have not been proposed for classification.

The robustness of GrAFX in complex activity segmentation, is measured with a dynamic time wrapping (DTW) based metric, which measures the consistency of the output under small changes in the hyper-parameters. Additionally, to analyze the stability of GrAFX in classification, we propose two metrics that quantify the variation in the loss function and the estimated probability of each class as different training sets are used to build the model. We

The work was funded by Mitsubishi Electric Corporation. Siheng Chen conducted this work while he worked at Mitsubishi Electric Research Labs.

demonstrate that our approach achieves better stability than state of the art systems, leading to improved generalization across datasets.

There are three main contributions in this paper. First, we introduce a spatio-temporal hand graph that can extract features that generalize across a range of applications, without requiring prior knowledge of the specific application. Second, we propose a DTW based measurement for unsupervised temporal action segmentation tasks, which can qualitatively analyze the performance across multiple subjects. Finally, we analyze the performance of our proposed recognition system based on two novel notions of stability: *validation loss stability* (LFS) and *estimation stability* (ESCV).

We start by introducing novel spatio-temporal hand graphs and studying their properties, including their computational efficiency. Then the graph-temporal transform of [18] is applied for feature extraction to graph signals defined on the constructed graphs. We provide an interpretation for the resulting representations, based on the spectrum and basis of the constructed graph, which help us justify their suitability for action recognition and segmentation tasks.

We present two sets of experiments. In *Experiment 1*, our goal is to segment an activity into small sub-tasks without having prior knowledge about the task. The stability of the complete system is also provided in terms of different choices of hyper-parameters, showing the advantages of an unsupervised system. In *Experiment 2*, we perform an action recognition experiment on the FPHA dataset [2] using only the hand position data. We provide a detailed stability analysis for our classification model using GrAFX and compare it to the state of the art LSTM model [19].

2. PROPOSED FEATURE EXTRACTION AND STABILITY METRICES

2.1. Hand graph selection

In this paper, we extend the fixed undirected graph representation of human hands of [12] to incorporate temporal graph connections. We consider both the finger-connected hand graph $\mathcal{G}_{FH}$ (21 nodes, 24 edges, see Fig. 1(a)) and the left-right hand graph $\mathcal{G}_{LRH}$ (42 nodes, 46 edges, see Fig. 1(b)). The elementary basis vectors corresponding to the spectrum of these spatial graphs can be used for feature extraction. As shown in Figs. 2 and 3, it is possible to interpret elementary basis vectors associated with different values of λ, the graph frequency. First, the basis vectors corresponding to the lowest frequency in both figures capture the average of the signal. The other two bases shown in Fig. 2, corresponding to medium frequencies, capture different types of intra hand motion. In Fig. 3, the basis corresponding to $\lambda = 0.12$ captures the variation between the thumb and the other fingers, $\lambda = 0.26$ corresponds to the variation in motion between hands, and $\lambda = 0.53$ shows intra hand motion variation.

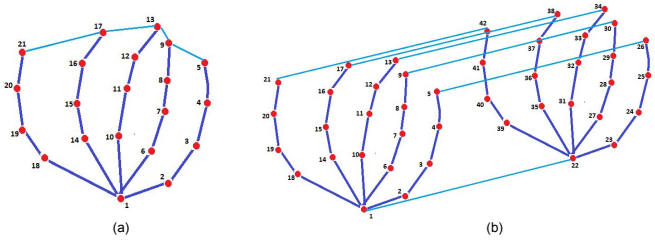

Fig. 1: (a) $\mathcal{G}_{FH}$ for single hand, (b)$\mathcal{G}_{LRH}$ for both hands.

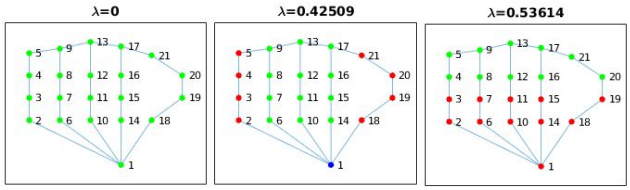

Fig. 2: Example elementary frequency basis for $\mathcal{G}_{FH}$. Green dot: positive value. Red dot: negative value. Blue dot: zero.

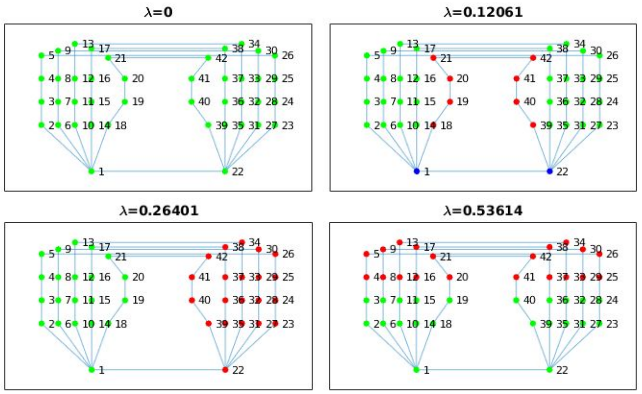

Fig. 3: Example elementary frequency basis for $\mathcal{G}_{LRH}$. Green dot: positive value. Red dot: negative value. Blue dot: zero.

2.2. Graph-based Application-agnostic Feature Extraction (GrAFX)

The input data is represented by a matrix $\mathbf{X}$, containing observations on each of the N joints for each of the T time instants. The observation at each joint contains the 3 dimensions of the corresponding motion vector at that joint. Each dimension is considered separately. Our proposed feature extraction is based on a separable transform as in [18], leading to the separable graph temporal Fourier transform (SGTFT):

$$F(\mathbf{X};\mathcal{G}) = \mathbf{\Psi}_{\mathcal{G}}\mathbf{X}\mathbf{\Psi}_{T} \tag{1}$$

where $\mathbf{\Psi}_T$ is a normalized discrete Fourier transform (DFT) matrix of size $T \times T$ and $\mathbf{\Psi}_{\mathcal{G}}$ is the $N \times N$ left eigenvector matrix of the Laplacian matrix $\mathbf{L}_G$ of $\mathcal{G}$. Using the properties of the Kronecker product ($\otimes$), SGTFTcan be written as $F(\mathbf{X};\mathcal{G}) = (\mathbf{\Psi}_{\mathcal{G}} \otimes \mathbf{\Psi}_T)\mathbf{X} = \mathbf{\Psi}_J\mathbf{X}$ where $\mathbf{\Psi}_J = \mathbf{\Psi}_{\mathcal{G}} \otimes \mathbf{\Psi}_T$. The Laplacian matrix of J can be expressed as $\mathbf{L}_J = \mathbf{I}_T \otimes \mathbf{L}_G \oplus \mathbf{L}_T \otimes \mathbf{I}_N$ so that:

$$\mathbf{L}_J = (\mathbf{\Phi}_T \otimes \mathbf{\Phi}_{\mathcal{G}})(\mathbf{\Lambda}_T \oplus \mathbf{\Lambda}_{\mathcal{G}})(\mathbf{\Psi}_T \otimes \mathbf{\Psi}_{\mathcal{G}}) = \mathbf{\Phi}_J\mathbf{\Lambda}_J\mathbf{\Psi}_J \tag{2}$$

where $\mathbf{\Phi}_T = \mathbf{\Psi}_T^{-1}$ and $\mathbf{\Phi}_{\mathcal{G}} = \mathbf{\Psi}_{\mathcal{G}}^{-1}$. The columns of $\mathbf{\Psi_J}$, denoted by $\boldsymbol{u}_k$, form a spectral basis for graph signals residing on $\mathcal{G}$, so that any $\boldsymbol{x}_i$ can be written as a unique linear combination of $\boldsymbol{u}_k$ as $\boldsymbol{x}_i = \sum_{k=1}^{N\times T} \alpha_{k,i}\boldsymbol{u}_k$, where

$$\alpha_{k,i} = \boldsymbol{x}_i^{\top}\boldsymbol{u}_k. \tag{3}$$

Let $\boldsymbol{x}_i$ be the motion vector present in each node (joint) of the graph (hand) and $\boldsymbol{\alpha}_k$ be a vector of length equal to the dimension of data at each joint (e.g., length three if motion at each joint is in 3D). Thus, $\alpha_{1,i}$, $\alpha_{2,i}$, ..., $\alpha_{N\times T,i}$ is a unique representation of the signal over a window of length T. Since these α features do not depend on a

specific application, we call them *graph-based application-agnostic features*. At any given time window, we form a graph signal where to each node of the hand graph we associate a motion vector with the corresponding motion of that joint. In this paper we apply GrAFX in two applications: unsupervised action segmentation and action recognition.

2.3. Unsupervised Activity Segmentation using GrAFX

2.3.1. Offline segmentation

For offline segmentation (*experiment1*), the hand skeleton data sequence is divided into smaller windows $\{w_1, w_2, ..., w_M\}$, where each window is represented by a feature vector computed using GrAFX . We use Expectation maximization (EM) [20] Gaussian mixture model based clustering for segmentation. Given the target number of actions, C, we define C clusters, and assume that the features of each cluster follow a different Gaussian mixture model. EM is then used to get the best set of parameters for those mixture models representing the subtasks.

In each iteration, the EM algorithm maximizes the likelihood and re-estimates the parameter values until it reaches the convergence criteria. After clustering, each window obtains a class label from the set $\{1, 2, ..., C\}$. When there is a change in class label, in between two consecutive windows w_t and w_{t+1}, we consider the end of window t as a good segmentation instance. Since each window is assigned to a cluster, the system is sensitive to the start and end of each window. To reduce the impact of window position, we use overlapping windows. However, this may result in two labels being assigned for for a given time interval. If an interval gets two different labels, the label with a higher likelihood is chosen as the current label.

2.3.2. Qualitative analysis of the performance

Consider a scenario where we do not have any prior knowledge about tasks performed by the subjects, but we have an example of those same tasks completed by an expert. Then, using the segmentation results (Section 2.3.1), we can rank how well each subject has completed the task by measuring the similarity of their work with that of the expert. To quantify the similarity between two unequal length data sequences of different subjects, we use dynamic time warping (DTW) [21] where $\text{dtw}(A_i, B_j)$ between sequence A with length i and sequence B with length j is defined as:

$$\text{dtw}(A_i, B_j) = \text{dist}(A_i, B_j) + \min(\text{dtw}(A_{i-1}, B_j), \text{dtw}(A_i, B_{j-1}), \text{dtw}(A_{i-1}, B_{j-1})) \quad (4)$$

We are given data sequences from an expert, ex (a reference sequence), and two subjects, s_1 and s_2, with respective lengths L_{ex} , L_{s1} , L_{s2}, where $L_{ex} < L_{s_1} < L_{s2}$. We compute $\text{dtw}(ex, s_1)$ and $\text{dtw}(ex, s_2)$ We use a normalized DTW so that a distance can be computed when $L_A \neq L_B$:

$$\text{dtw}_{\text{norm}}(A_{L_A}, B_{L_B}) = \frac{\text{dtw}(A_{L_A}, B_{L_B})}{\max(L_A, L_B)} \quad (5)$$

To compare $\text{dtw}_{\text{norm}}(ex, s_i)$ across subjects (s_i), we apply *min-max* normalization [22] to each sequence, which removes biases in the range of values in the data.

2.4. Activity Recognition using GrAFX

2.4.1. Proposed Stability Metrics for Activity Recognition

Activity recognition is performed using GrAFX to extract spatio-temporal features from the hand skeleton data. A temporal window is defined to capture the local temporal variation of the data, followed by a mean pulling algorithm to obtain a 1D feature vector per window. Then SVM [23] is used for classification. Stability analysis can be used to determine the sensitivity of the output of the system to input variations. In this paper, we use three different measures of stability.

First, *leave-one-out cross-validation stability* (LOOCS) is defined as:

$$\psi_{LOOCS} = \frac{\sigma_{acc}}{|\mu_{acc}|} \quad (6)$$

where μ_{acc} and σ_{acc} denote the mean and standard deviation of the accuracy over different test settings. Here, ψ_{LOOCS} is computed in a cross-subject setting, i.e., we train with all subjects but one, and test on the subject we left out.

We also quantify *Loss function stability* (LFS). While training any classification model, the main goal is minimizing the loss function while tuning the parameters. So, it is crucial to observe the variation in the loss function for a specific validation set when the training set is changed. To quantify this, we measure validation loss stability for each validation set and compute the average. Assume there are M validation sets $\{vs_1, vs_2, ..., vs_M\}$ and for each validation set there are N training sets $\{ts_1^i, ts_2^i, ..., ts_N^i\}$ where i denotes the corresponding validation set. For each validation set, we compute N validation losses $\{VL_1^i, VL_2^i, ..., VL_N^i\}$. Let μ_{VL^i} and σ_{VL^i} denote the mean and standard deviation of the validation loss value over different test settings. Thus we define LFS, ψ_{LFS}, as:

$$\psi_{\text{LFS}} = \frac{1}{M} \sum_{i}^{M} \frac{\sigma_{VL^i}}{|\mu_{VL^i}|} \quad (7)$$

While computing the validation loss (e.g., cross-entropy loss in LSTM or hinge loss in SVM), we only consider the estimated score or the probability of the true class, but for a stable model, the probabilities of the other classes should be close as well for a particular validation set and varying training set. Inspired by the concept of *estimation stability* [17], we introduce *estimation stability* (ESCV). Let the predicted probabilities of a model for each class be denoted by p_i^c, where c and i denote the class index and validation sample of i^{th} validation set respectively. For each class, we compute ψ^c by taking the ratio of the standard deviation and the absolute mean of the class probability. Finally, ESCV, ψ_{ESCV}, is the average over all the class and different validation set.

$$\psi_{\text{ESCV}} = \frac{1}{MN} \sum_{c=1}^{N} \sum_{i=1}^{M} \psi_i^c \quad (8)$$

3. EXPERIMENTAL RESULTS

3.1. Datasets

USC Toy Assembling Dataset: Das et al. [12] introduce a toy assembling dataset, where each subject assembles a GoPiGo3 [1] robot base kit according to a specific set of instructions. This task had three main sub-actions: assembling, combining, and checking. 11

[1] https://www.dexterindustries.com/store/gopigo3-base-kit/

Table 1: Segmentation Accuracy for USC Dataset [12]

Features	Motion Vectors	$\mathcal{G}_H$ [12]	GrAFX
seg_{acc}	62.15%	79.58%	**86.12%**

subjects participated in the task and each performed the task three times. The 2D hand skeleton key points for each subject were extracted by OpenPose at 30 frames per second.

First-Person Hand Action Benchmark: Garcia et al. [2] proposed a first-person dynamic hand actions (FPHA) where subjects interact with 3D objects. This dataset contains 1,175 action videos belonging to 45 different action categories involving 26 different objects, in 3 different scenarios, kitchen, office, and social, and performed by 6 subjects. In our experiments, we only use the right-hand pose annotations data.

3.2. Experiment 1: Segmentation Results

Fig 4 shows the similarity between each subject and the expert, where iteration 1 for subject 11 was used as the expert reference ($S_{ex} = s_{11}, iter_1$). We compute dtw$_{norm}$ between each data sequence and S_{ex}. Lower values of dtw$_{norm}$ correspond to greater similarity to the expert, implying better performance. Note that each subject performed the task three times, and it can be seen that their performance improved with each attempt, as shown in Fig. 4, with 10 out of 11 subjects achieving their best results in the third attempt.

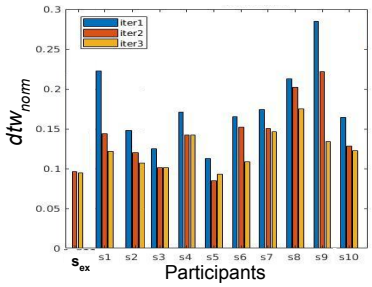

Fig. 4: dtw$_{norm}$ for each data sequence considering s_{11} as expert.

Let segmentation accuracy (seg_{acc}) be defined as the ratio of number of windows clustered correctly to the total number of windows. Table 1 summarizes seg_{acc} computed using different features extraction methods. It is clearly seen that GrAFX performs the best.

To evaluate the robustness of our algorithm, we repeat the similarity-based ranking of subjects under different conditions where we vary parameters such as window-size (chosen with values 2/3/4/5/7 secs) and number of subtasks. After segmenting each data sequence, the distance from the expert is computed using (5), then we select the top ranking user, i.e., the one with smallest distance to the expert data. Out of 33 subjects, the same subjects were in the top 3 ranking 46% of time for varying window size, and 76% of time when considering top 15 rankings.

3.3. Experiment 2 : Recognition results

To evaluate the performance of the hand graph in an action recognition task, we use the FPHA dataset. To compare our results with the LSTM based action recognition mentioned in [2], we use similar experimental settings with train:test dataset ratios of 1:3, 1:1 and 3:1 at

Table 2: Recognition accuracy of FPHA dataset

Protocol	LSTM [2]	$\mathcal{G}_H$ [12]	GrAFX
1:3	58.75	48.63	62.39
1:1	78.73	63.3	79.14
3:1	84.82	72.46	84.93
Cross-person	62.06	61.55	**73.67**

Table 3: Recognition accuracy of FPHA in train:test=1:1 protocol

Algorithm	Accuracy	Algorithm	Accuracy
JOULE-pose [24]	74.60	Gram Matrix [25]	85.39
HBRNN [26]	77.40	TCN-16 [27]	76.28
TF [28]	80.69	TCN-16 + TTN [11]	80.14
Lie Group [29]	82.69	$\mathcal{G}_{tH}$	**79.14**

sequence level. The second protocol consists of a 6-fold 'leave-one-person-out' cross-validation, i.e., each fold consists of 5 subjects for training and 1 for testing. In this set, the training and testing set do not share the same subjects taking care of and subjective bias.

As we can see clearly from the Table 2 GrAFX with spatio-temporal connection outperforms $\mathcal{G}_H$ [12] with only spatial connection in terms of accuracy. While comparing with the LSTM, our proposed method outperformed by 5% for the cross-person setting, which is a more realistic scenario in practice. GrAFX shows comparable performance to the state of the art method in a 1:1 train test setting as shown in Table 3. Though the feature extraction of GrAFX is unsupervised, the proposed features uniquely capture the action characteristics rather than any person specific information. In contrast, while LSTM, being a data-driven method, performs poorly in the cross person setting.

Table 4 summarizes the stability analysis (Section 2.4) of these two classifications models. Though the accuracy of the proposed model and LSTM is comparable in some cases, GrAFX outperforms LSTM in cross-person setting both in accuracy and stability.

Table 4: Stability analysis of the classification models

Method	ψ_{LOOCS}	ψ_{LFS}	ψ_{ESCV}
LSTM	0.106	0.129	1.106
GrAFX	0.0267	0.0455	0.0372

As an additional comparison, we use principal component analysis [30], a data driven approach, to extract features followed by a SVM [23] for classification. This method achieves comparable performance in stability with the proposed method, but lacks in terms of performance, with an average accuracy of 45%.

4. CONCLUSION

In this paper, we propose a spatio-temporal hand graph-based application-agnostic feature extraction method for hand motion analysis. To evaluate the performance of this method, we choose two complex activities, an unsupervised segmentation task and a supervised classification task. We introduce DTW based stability metric to measure the robustness of the segmentation algorithm and LOOCS, LFS, ESCVto analyze the robustness of GrAFX in classification. GrAFX achieves better stability as compared to the state-of-the-art algorithms.

5. REFERENCES

[1] T Han, J Wang, A Cherian, and S Gould, "Human action forecasting by learning task grammars," *arXiv:1709.06391*, 2017.

[2] G Garcia-Hernando, S Yuan, S Baek, and T K Kim, "First-person hand action benchmark with rgb-d videos and 3D hand pose annotations," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 409–419.

[3] V Srinivasan, S Lapuschkin, C Hellge, K R Müller, and W Samek, "Interpretable human action recognition in compressed domain," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 1692–1696.

[4] P. Bharti, D. De, S. Chellappan, and S. K. Das, "HuMAn: Complex activity recognition with multi-modal multi-positional body sensing," *IEEE Transactions on Mobile Computing*, vol. 18, no. 4, pp. 857–870, April 2019.

[5] Z Cao, T Simon, S E Wei, and Y Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 7291–7299.

[6] W Zhu, C Lan, J Xing, W Zeng, Y Li, L Shen, and X Xie, "Co-occurrence feature learning for skeleton based action recognition using regularized deep lstm networks," in *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[7] S Yan, Y Xiong, and D Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[8] T Kerola, N Inoue, and K Shinoda, "Spectral graph skeletons for 3d action recognition," in *Asian Conference on Computer Vision*. Springer, 2014, pp. 417–432.

[9] J Y Kao, A Ortega, D Tian, H Mansour, and A Vetro, "Graph based skeleton modeling for human activity analysis," in *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019, pp. 2025–2029.

[10] M Li, S Chen, X Chen, Y Zhang, Y Wang, and Q Tian, "Actional-structural graph convolutional networks for skeleton-based action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3595–3603.

[11] S Lohit, Q Wang, and P Turaga, "Temporal transformer networks: Joint learning of invariant and discriminative time warping," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12426–12435.

[12] P Das, J Y Kao, A Ortega, T Sawada, H Mansour, A Vetro, and A Minezawa, "Hand graph representations for unsupervised segmentation of complex activities," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 4075–4079.

[13] Quentin De S, H Wannous, and J P Vandeborre, "Heterogeneous hand gesture recognition using 3D dynamic skeletal data," *Computer Vision and Image Understanding*, vol. 181, pp. 60–72, 2019.

[14] H. Wang, T. M. Khoshgoftaaar, and Q. Liang, "Stability and classification performance of feature selection techniques," in *2011 10th International Conference on Machine Learning and Applications and Workshops*, Dec 2011, vol. 1, pp. 151–156.

[15] F Bonassi, E Terzi, M Farina, and R Scattolini, "Lstm neural networks: Input to state stability and probabilistic safety verification," in *Learning for Dynamics and Control*. PMLR, 2020, pp. 85–94.

[16] O Bousquet and A Elisseeff, "Stability and generalization," *Journal of machine learning research*, vol. 2, no. Mar, pp. 499–526, 2002.

[17] C Lim and B Yu, "Estimation stability with cross-validation (ESCV)," *Journal of Computational and Graphical Statistics*, vol. 25, no. 2, pp. 464–492, 2016.

[18] A Loukas and D Foucard, "Frequency analysis of time-varying graph signals," in *2016 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*. IEEE, 2016, pp. 346–350.

[19] S Hochreiter and journal=Neural computation volume=9 number=8 pages=1735–1780 year=1997 publisher=MIT Press Schmidhuber, J, "Long short-term memory," .

[20] L. Xu and M. I. Jordan, "On convergence properties of the em algorithm for gaussian mixtures," *Neural computation*, vol. 8, no. 1, pp. 129–151, 1996.

[21] S Masood, M P Qureshi, M B Shah, Salman Ashraf, Zahid H, and G Abbas, "Dynamic time wrapping based gesture recognition," in *2014 International Conference on Robotics and Emerging Allied Technologies in Engineering (iCREATE)*. IEEE, 2014, pp. 205–210.

[22] S Patro and K K Sahu, "Normalization: A preprocessing stage," *arXiv preprint arXiv:1503.06462*, 2015.

[23] B Scholkopf and A J Smola, *Learning with kernels: support vector machines, regularization, optimization, and beyond*, MIT press, 2001.

[24] J F Hu, W S Zheng, J Lai, and J Zhang, "Jointly learning heterogeneous features for rgb-d activity recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 5344–5352.

[25] X Zhang, Y Wang, M Gou, M Sznaier, and O Camps, "Efficient temporal sequence comparison and classification using gram matrix embeddings on a riemannian manifold," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4498–4507.

[26] Y Du, W Wang, and L Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1110–1118.

[27] T. S. Kim and A. Reiter, "Interpretable 3D human action analysis with temporal convolutional networks," in *2017 IEEE conference on computer vision and pattern recognition workshops (CVPRW)*. IEEE, 2017, pp. 1623–1631.

[28] G Garcia-Hernando and T K Kim, "Transition forests: Learning discriminative temporal transitions for action recognition and detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 432–440.

[29] R Vemulapalli, F Arrate, and R Chellappa, "Human action recognition by representing 3D skeletons as points in a lie group," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 588–595.

[30] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics and intelligent laboratory systems*, vol. 2, no. 1-3, pp. 37–52, 1987.