Efficient Dense-Graph Convolutional Network with Inductive Prior Augmentations for Unsupervised Micro-Gesture Recognition

Atif Shah, Haoyu Chen, Henglin Shi and Guoying Zhao*
Center for Machine Vision and Signal Analysis (CMVS),
University of Oulu, Finland
Email: [atif.shah, chen.haoyu, henglin.shi, guoying.zhao]@oulu.fi

***Abstract*—Skeleton-based action/gesture recognition has already witnessed excellent progress on processing large-scale, laboratory-based datasets with pre-defined skeleton joint topology. However, it's still an unsolved task when it comes to real-world scenarios with practical limitations such as small-scaled dataset sizes, few-labeled samples, and various skeleton topologies. In this paper, we worked on the recognition of micro-gesture, which are subtle body gestures collected in real-world scenarios. Specifically, we utilize contrastive learning to heritage the knowledge from known large-scale datasets for enhancing the learning on fewer samples of micro-gestures. To overcome the gap caused by various domain distribution and structure topology between the datasets, we compute skeleton representations from augmented sequences via momentum-based efficient and scalable encoders as additional inductive priors. Importantly, we proposed an effective dense-graph based unsupervised architecture that resorts to a queue-based dictionary to store positive and negative keys for better contrast with queries to learn substantially efficient and discriminant patterns in the feature space. Together with cross-dataset experimental results show that our model significantly improves the accuracies on two micro-gesture datasets, SMG by 7.4% and iMiGUE by 18.41% advocating its superiority.**

I. Introduction

Human action/gesture analysis is an important area of research in computer vision where the goal is to automatically detect, recognise and interpret human behaviors from the data available as either images, videos or skeleton. For this purpose, a large amount of RGB data has been used. Due to the recent developments of depth sensors [1], the task of action/gesture recognition is revolutionized and the community has witnessed a shift towards depth images as well as skeleton data. The skeleton data is a compact representation of the human body where the human body joints are represented in three-dimensional coordinates as graph data. Leveraging from such graph data of the joint coordinates, 3D human skeleton modeling has proved to be more effective by utilizing various positions and viewpoints [2].

Most of the existing skeleton-based methods use supervised learning methods which rely on a handsome amount of annotated data [3], [4], [5]. However, labeling such large datasets, especially in the wild scenarios, is a tedious job and needs a workforce thereby making the overall process time-expensive. Importantly, the inevitable data biasedness due to human involvement and higher inter-class similarities makes the data labeling task more difficult, ultimately leading to mislabeling or uncertain labeling [6]. Besides, small-scaled datasets sizes, few labeled samples, and various skeletons are all unsolved issues in practice of real-world scenarios, unsupervised methods provide effective ways to learn the gesture representations from data without the need of manual annotations.

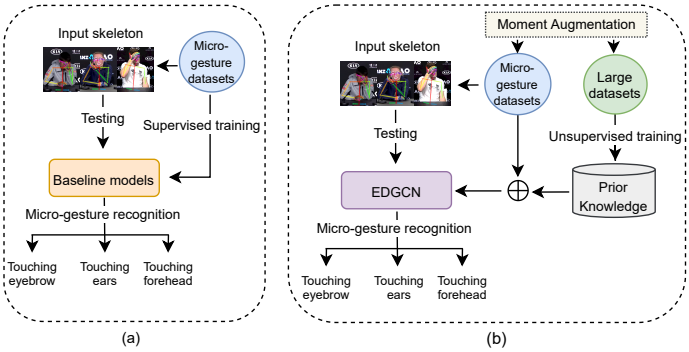

Fig. 1. An illustration of supervised and unsupervised methods (a) shows the supervised model for MG dataset, (b) shows the proposed unsupervised method with inductive prior moment-based augmentation and pre-trained model knowledge. MG samples are borrowed from iMiGUE dataset.

In gesture recognition, a challenging task is recognising Micro-Gestures (MG) which are different than normal gestures. Normal gestures are used to convey certain messages or feelings directly such as waving hands for hello or goodbye. However, MG are unintentional behaviors induced by a person's inner feelings, such as rubbing hands because of stress. Such gestures are involuntary emotional actions that exhibit a person's true emotions. MG are subtle, swift and used to hide a person's internal emotions, making it more challenging task than normal gestures and action recognition [7], [8].

Keeping the aforementioned challenges in mind, in this work, we propose an unsupervised learning framework to recognise MG from two distorted versions of the incoming batches of data which are firstly augmented. The augmented data is then fed to the model's core, where a couple of

* indicates corresponding author

dense graph-based encoders are used to learn from augmented instances of the skeleton sequences. To maximize the similarity between the query and positive keys for better gestures representation, contrastive learning is used. Following on, the learned queries are finally fed to a linear classifier to perform MG recognition. An illustration of our proposed unsupervised learning framework is shown in Figure 1.

In the following, we highlight several salient features of our work.

- We introduce a novel and efficient extension of the existing MSG3D model, namely Efficient Dense-Graph Convolutional Network (EDGCN).
- To our knowledge its the first attempt to utilize the prior knowledge from large datasets with unsupervised learning to improve the performance of small real-world datasets for skeleton based MG recognition.
- Another upshot of our proposed method is that it is equipped with the moment-based feature augmentation strategy that substantially reduces the knowledge-gap between large and small MG datasets.
- Our efficient model design additionally offers a robust framework to yield the-state-of-art results on MG datasets with both intra- and cross-dataset settings.

The rest of the paper is organized as follows: Section II discusses the related work in the literature, while Section III explains the methodology of our unsupervised model. Section IV presents the experimental setups, results, and discussions, whereas Section V concludes the paper.

II. Related Work

The action/gesture recognition methods are divided into two main categories, supervised methods and unsupervised methods, both are reviewed in this section.

A. Skeleton-Based Supervised Methods

A skeleton-based data consists of 2D and 3D points of joints of human body. It has some benefits over images and videos because it is more robust against noises brought by such as lighting conditions and background cluttering, and could focus more on gestures and actions. Yan *et al.* [3] proposed a spatial-temporal graph convolutional network (STGCN) to extract complex information from human skeleton for action recognition. Si *et al.* [4] exploited Long Short-Term Memory (LSTM) along with convolutional networks to improve the performance using better discriminatory spatial and temporal features. Shi *et al.* [5] used a two-stream adaptive graph convolutional network (2s-AGCN) which has a data-dependent graph with first and second-order bone information for action recognition. Liu *et al.* [9] introduced the multiscale unified spatial-temporal graph convolutional operator (MS-G3D). Unlike ST-GCN, a single node is connected to its neighbors in temporal space which creates a dense connection to the next temporal nodes.

However, all the aforementioned methods need a large amount of annotated data to fully preserve the feature representation of the target data distribution and failed to efficiently learn the feature representation with small-scale datasets. Instead, we propose a novel and efficient dense-graph network, the architecture of which is optimized for unsupervised learning with fewer labeled samples.

B. Skeleton-Based Unsupervised Methods

Recent supervised approaches have achieved robust performance, however, unsupervised setup is more advantageous, where is no sufficient data with reliable annotation. Zheng *et al.* [2] used unsupervised representation learning to capture global motion dynamics. The generative adversarial network (GAN) is used as encoder-decoder for modeling motion dynamics and learning discriminative features for recognising actions. Similarly, the method proposed by Su *et al.* [10] is benefited from encoder-decoder recurrent neural network (RNN) to learn features for action recognition. Lin *et al.* [11] integrated jigsaw puzzle, motion prediction and contrastive learning to learn more generalized representation and used unsupervised Bidirectional-Gated Recurrent Unit (Bi-GRU) encoder for action recognition. Li *et al.* [12] combined RGB and depth images and used unlabeled data to learn view-invariant action and predict action 3D motion. One of the approaches to achieve unsupervised learning is contrastive learning [13].

Contrastive learning is an effective unsupervised learning paradigm which has been used for various pretext tasks [14]. One of the main contributors to contrastive learning is contrastive loss which represents the similarity among pairs in a space. The noise-contrastive estimation (NCE) [15] pulls the similar instances of augmented samples and pushes away the distinct ones. The contrastive multi-view coding (CMC) [16] maximizes mutual information among different views, and momentum contrastive paradigm (MoCo) [17] contributes to the learning by using momentum-based updates and queue-based dictionary.

C. Skeleton-based micro-gestures datasets

Most of the literature focused on laboratory collected dataset [18], [19] for the skeleton-based action recognition task. Those datasets are large-scale, category balanced, multi-view collected, which are idealized settings for machine learning methods. However, for in-the-wild datasets [7], [8], [20] based on real-world scenarios, classes are commonly unbalanced and the samples size is smaller as compared to laboratory collected datasets, which makes the task more challenging. Some datasets used actors to perform certain tasks, however, in a real-world scenario, even for same action, there would be many different styles when performed by different people. Skeleton contains topological and geometrical properties of shapes, such as length, width, direction and connectivity. Different datasets use various skeleton topologies, such as NTU RBG+D 60 [18] and SMG [8] dataset use Kinect (Kinect v2) while the iMiGUE [7] datasets uses OpenPose [21] skeleton topology. This makes the action/gesture recognition task more challenging for models to learn features from different topologies [22].

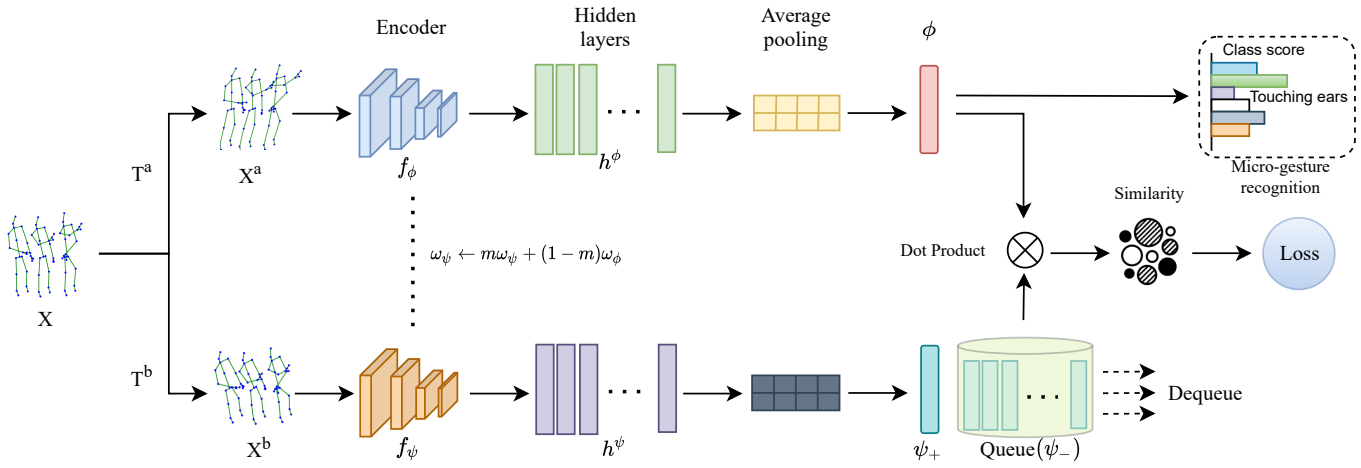

Fig. 2. An overview of our proposed dense-graph based unsupervised learning framework for efficient MG recognition.

III. Methodology

In this work, we propose an unsupervised-based model to capturing MG features (representations) without providing the labeled data. The model takes a skeleton sequence as input, converts it into two random augmentations and feeds them to the encoders. We use EDGCN (query) and momentum-based EDGCN (key) encoders (discussed in Section III-B). The encoders take the augmented data into hidden layers followed by the pooling layers where the temporal average pooling is applied. The query encoder updates their weights via backpropagation while the key encoder optimizes its weights through the momentum-based update. Till then we have the augmented instances in the keys and query where the contrastive learning takes place for better gestures representation. The output of the query encoder is then fed to a linear classifier with their original labels to perform the final step which is MG recognition as shown in Figure 2. The proposed model is evaluated through two MG datasets which are SMG and iMiGUE and action recognition dataset NTU RGB+D 60 for cross-evaluation.

A. Unsupervised Framework

An input sequence X, consists of T consecutive skeleton frames $X = \{x_1, .., x_T\}$ where $X \epsilon \mathbb{R}^{M \times J \times 3}$, M is the number of persons, J is number of joints and 3 corresponds to the dimensions. The training set is $\zeta = \{X^i\}_{i=1}^N$ where N is the number of skeleton sequence acquired from multiple persons. Each sequence x^i corresponds to label Y_i, where $Y_i \epsilon \{c_1, ..., c_z\}$ and z corresponds to gesture classes. We need to learn gesture representation ϕ effectively from the sequence X without providing the label Y_i. The learned representation ϕ is then evaluated by linear protocol and fed to a linear classifier with labels to classify MG.

Several augmentation strategies were used to learn more robust and discriminatory features from unlabeled sequences. These augmentation strategies include rotation, shear, reverse, Gaussian noise, Gaussian blur, joint mask and channel mask borrowed from [23].

Two encoders are used to learn discriminatory features from the augmented instances of the sequence without providing the sequence labels. While the query encoder f_ϕ learns and updates their weights through backpropagation, the key encoder f_ψ updates their weights with the collaboration of query encoder f_ϕ using weighted average, such as momentum average. The f_ψ encoder followed by two keys, ψ_+ represents the positive keys that contains the current batch keys, whereas the queue contains negative keys ψ_- from immediate batches, therefore it is hard for a model to update parameters with a large number of keys in a dictionary which is intractable [17]. During the training phase, the f_ψ model updates its weights using following equation:

$$\omega_\psi \leftarrow m\omega_\psi + (1-m)\omega_\phi \tag{1}$$

where ω_ψ and ω_ϕ are parameters of f_ψ and f_ϕ encoder and m is the update rate which lies between 0 and 1.

The augmented sequence X^a and X^b encoded by the hidden layer of the encoder as follows:

$$h^\phi = f_\phi(X) = \sum_{k=0}^{K} D_{(\tau,k)}^{-1/2} A_{(\tau,k)} D_{(\tau,k)}^{-1/2} [X_{(\tau)}^{a(l)}] t\Theta_k^l \tag{2}$$

$$h^\psi = f_\psi(X) = \sum_{k=0}^{K} D_{(\tau,k)}^{-1/2} A_{(\tau,k)} D_{(\tau,k)}^{-1/2} [X_{(\tau)}^{b(l)}] t\Theta_k^l \tag{3}$$

where A represents the adjacency matrix, D is a diagonal degree matrix, Θ^l shows the learnable weights of matrix at layer l, X represents the features and t shows the time, while τ and k are the window size and scales of aggregation, respectively [9].

After hidden layers, the average pooling is performed on all the hidden layer across the temporal domain of each encoder, f_ϕ and f_ψ.

$$\phi = \frac{1}{T}\sum_{i=1}^{T} h_{\phi_i} \tag{4}$$

$$\psi_+ = \frac{1}{T}\sum_{i=1}^{T} h_{\psi_i} \tag{5}$$

where ψ_+ represents the positive keys in an augmented sequence. In every batch, the positive keys correspond to ϕ to learn the discriminatory features via contrastive learning.

A queue-based dictionary is used for storing the positive keys ψ_+ in each mini-batch from the augmented sequence. The previous batch keys are considered as negative keys ψ_- in a queue and the oldest ψ_- is dequeued to keep the dictionary up-to-date and consistent with query ψ for better contrastive learning.

Our goal is to learn more discriminatory features and accurately classify gestures in a sequence therefore we need a better loss function that can maximize the similarity between query ϕ and all the keys ψ while reducing the loss. The loss function we used is as follows:

$$loss = -log\frac{exp(\phi.\psi_+/\lambda)}{exp(\phi.\psi_+/\lambda) + \sum_{i=1}^{K} exp(\phi.\psi_-^i/\lambda)} \tag{6}$$

where λ controls the learning speed and K is the number of negative keys ψ_-in a queue.

After all the aforementioned process, the query ϕ has been learned effectively through contrasting with keys, now the query ϕ will be fed to the linear classifier with their original labels for training and will be kept frozen during linear evaluation.

B. Efficient dense-graph convolutional network

Inspired by the huge success of the MSG3D [9] model in the task of supervised action recognition, we propose to heritage the MSG3D architecture as an encoder because it captures complex joint correlation across the spatial and temporal domain. However, we find that the original MSG3D model for unsupervised learning tasks is too complex to learn discriminate features. Since it learns from the dense constructed graphs of skeleton topologies, too detailed graph structures will let it suffer from the overfitting issue in the contrastive learning (see later experimental parts). To this end, we propose an efficient dense-graph convolutional network (EDGCN). The EDGCN model learns features across spacetime with less number of aggregations and small range modeling instead of a large number of aggregation and long-range modeling. Specifically, instead of spanning the feature channel into multi-scales to construct hyper-dense graphs, we restricted the input feature channels to a minimum scale without hurting the effectiveness of computing complex correlations by utilizing the strength of moment augmentation (see Section III-C) fix the convolutional window in temporal modeling and restricted it to a minimum scale without hurting the effectiveness of computing complex correlations. Besides, instead of constructing multiple scales of temporal channels and aggregating multiples of branches, we implement simple temporal convolutional layers to the end of each aggregation of the spatial features from the graph convolutional layers (gcn3d1, gcn3d2 and gcn3d1 layers). In this way, the redundant channels of the dense graph are significantly pruned for unsupervised learning while rich spatial and temporal information is still preserved.

C. Moment-Based Augmentation

Data augmentation is useful to improve the model performance and enhance model generalization. In this paper, we used the moment-based augmentation strategy proposed in [24], The intuition of moment-based augmentation is that, to fully learn the detailed information of graph topology, we need to construct dense convolutional network for better non-linear representing abilities. However, we argue that, by augmenting the input skeleton joints from linear space into non-linear space with moment representations, it can be beneficial to the procedure of gradient descent and network optimization. Specifically, given a 3D point (x, y, z), we made use of its coordinates x, y, z and the products of them, namely $x^2, y^2, z^2, xy, xz, yz$ as its feature vector. In fact, from the point of view of moments and moment invariants, the features x, y, z can be regarded as the first-order geometric moments of the point, and $x^2, y^2, z^2, xy, xz, yz$ are its six second-order moments, and so on.

IV. Experiments and Results

We conducted various experiments with the following real-world MG datasets.

A. Datasets

1) Real-world micro-gesture datasets: We use two real-world datasets that are collected from in-the-wild scenarios.

Spontaneous Micro-Gesture (SMG) dataset [8] consists of 3,692 samples 17 MG. The datasets were collected from 40 subjects during narrating a fake and real story with 25 3D joints collected by Kinect.

Micro-Gesture Understanding and Emotion analysis (iMiGUE) dataset [7] consists of 32 MG collected from post-match press conferences videos. The training set consists of 13,936 and a testing set of 4,563 samples of MG to detect negative and positive emotions with 25 3D joints collected from OpenPose. Note that, following the protocol of [3], [9], we regard the 3rd dimensional of the OpenPose joints (the estimated probability of the joints) as a pseudo dimension.

2) Action Recognition dataset: **NTU RGB+D 60** (NTU-60) [18] consist of 56,578 skeleton sequences with 60 labels. The data was collected from 40 individuals and captured by three different camera angle views. Each frame consists of 1 to 2 subjects with N skeleton points, where N=25 body joint nodes with their 3D locations. This dataset follows two protocols, however, we use the first protocol for experiments: 1) Cross-subject (X-Sub) where 40 subjects are splits into 40,091 training and 16,487 testing samples. 2) Cross-view (X-View) where 18,932 samples from camera 1 used for testing while the rest of 37,646 samples from another cameras for training.

TABLE I
ABLATION STUDY RESULTS ON SMG DATASET

Network	Configuration	Results (%)	Model parameters
MSG3D [9]	Full model (13, 6)	37.2	4.36m
MSG3D [9]	Simplified model (6, 3)	44.6	2.46m
EDGCN	(1, 1)	**46.9**	**1.9m**

*m = million

TABLE II
INTRA-DATASET RESULTS AND COMPARISONS WITH TOP-1 ACCURACY(%)

Dataset	Methods	Top-1
SMG	MSG3D [9]	44.6
	EDGCN (ours)	46.9
	A-EDGCN (ours)	**47.9**
iMiGUE	P&C (Encoder-Decoder) [10]	31.67
	U-S-VAE Z (Encoder-Decoder) [7]	32.43
	MSG3D [9]	36.9
	EDGCN (ours)	36.8
	A-EDGCN (ours)	**37.5**

B. Implementation Details

The proposed method uses EDGCN as query encoder ϕ and momentum-based EDGCN as key encoder ψ with unlabeled data to recognise gestures. We use an SGD optimizer with a learning rate of $1e^{-5}$, the momentum of 0.9 and the weight decay of $1e^{-4}$. We set the contrastive learning rate λ to 0.06 and queue size to 500. We use 60 epochs for unsupervised gestures representation learning and 90 epochs for linear evaluation. In the linear evaluation step, we use a single-layer linear classifier with frozen parameters and use SGD optimizer with a momentum of 0.9 and weight decay of 0.5 multiplied at 15, 35, 60 and 75 epochs.

We evaluate and compare the results in two different settings, Intra-dataset and Cross-dataset. In Intra-dataset setting, we use three different models, to evaluate the datasets. 1) MSG3D [9] method in unsupervised manners; 2) EDGCN: Our proposed method and 3) Augmented EDGCN (A-EDGCN) used to enhance the performance of our proposed method via moment augmentation. We also compare our proposed methods to previously published methods, such as P&C [10] and U-S-VAE Z [7].

In Cross-dataset setting, we evaluate SMG and iMiGUE as a target datasets with pre-trained models on SMG, iMiGUE and NTU-60 datasets.

C. Ablation study

We conduct the ablation study of EDGCN on SMG dataset to prove the efficiency of its architecture. Table I shows the results of ablation study. The second column shows the number of scales of GCN and G3D, respectively, which is one of the hyper-parameters used in the MSG3D model [9]. Tabel I shows the unsupervised MSG3D models that consist of a complex graphical structure to extract features from real-world datasets leads to model overfitting. The real-world datasets consist of fewer samples and unbalanced classes, which makes the feature learning process more challenging. The first row of Tabel I shows the MSG3D model that achieves the accuracy of 37.2% with the 4.36 million model parameters. To improve the learning, we reduce the model scales significantly in simplified MSG3D (second row) and empower the model to learn more discriminatory features rather than learn redundant features. The simplified MSG3D improves the accuracy by 7.2 which reaches 44.6% and reduces the model parameters by 43%. Lastly, our proposed method EDGCN performs better than the other methods with an improvement of 2.3 in accuracy that reached 46.9% and further reduces the model parameters by more than 20% as shown in the last row of Tabel I.

TABLE III
CROSS-DATASET RESULTS AND COMPARISONS WITH TOP-1 ACCURACY(%)

Target dataset	Skeleton topology	Target dataset samples	Pre-trained dataset	Pre-trained dataset samples	Skeleton topology	Top-1
SMG	Kinect	3,692	Pre-trained SMG	3,692	Kinect	46.9
					Kinect (MA)	**47.9**
			Pre-trained NTU	56,578	Kinect	48.4
					Kinect (MA)	40.2
			Pre-trained iMiGUE	18,499	Openpose	39.2
					Openpose (MA)	**39.8**
iMiGUE	Openpose	18,499	Pre-trained iMiGUE	18,499	Openpose	36.8
					Openpose (MA)	**37.5**
			Pre-trained NTU	56,578	Kinect	34
					Kinect (MA)	**34.9**
			Pre-trained SMG	3,692	Kinect	35.3
					Kinect (MA)	34.9

*MA = Moment Augmented

D. Intra-dataset comparisons

We compare the results on the SMG dataset with the unsupervised MSG3D baseline [9] which achieve the accuracy of 44.6% as shown in first row of Table II. Our proposed EDGCN model performs better with 46.9% accuracy. With moment-based augmentation, the A-EDGCN model outperforms the aforementioned models with an accuracy of 47.9%.

Similarly, for iMiGUE dataset, by using the code of P&C [10] method, we achieved an accuracy of 31.67% and our baseline MSG3D achieve 36.9% accuracy which outperforms the previous method U-S-VAE Z [7] with accuracy 32.43% as shown in row three and four of Table II.

The unsupervised MSG3D on iMiGUE dataset performs comparatively to our EDGCN model (36.9% vs 36.8%). However, the moment-based augmented A-EDGCN model outperforms the other methods and achieves the accuracy of 37.5% as shown in last row of Table II. This means that our model improves the accuracy on SMG by 7.4% (44.6% vs 47.9%) and the accuracy on iMiGUE by 18.41% (31.67% vs 37.5%) from their initial baselines which is a substantial improvement for an unsupervised model.

E. Cross-dataset comparisons

Further experiments are also conducted via cross-dataset to benchmark our model performance. In this set of experiments, we use pre-trained models of three datasets, SMG, iMiGUE and NTU-60. The SMG and NTU-60 datasets have

the same Kinect skeleton topology while the iMiGUE skeleton topology is OpenPose. We cross evaluate the SMG dataset using pre-trained models from NTU-60 and iMiGUE and achieved the accuracy of 48.4% and 39.2%, respectively. However, using moment-based augmentation, the accuracy of pre-trained iMiGUE model is improved to 39.8% from 39.2% meanwhile the moment-based augmented pre-trained NTU-60 model didn't outperform the unaugmented pre-trained NTU-60. One of the reason that pre-trained model on NTU-60 performs slightly better than the pre-trained on SMG and iMiGUE models is that the NTU-60 has a large number of training samples which performance also reflects in the current experiments as shown in Table III.

On the other hand, we use similar pre-trained models to evaluate the iMiGUE dataset. Table III shows that pre-trained NTU-60 and pre-trained SMG did not perform well and got lower accuracy than the pre-trained iMiGUE dataset. One of the reasons is both NTU-60 and SMG have the same skeleton topology and are different than iMiGUE topology which affects the performance. However, still the moment augmented pre-trained NTU-60 model improves the accuracy by 2.35% than its unaugmented NTU-60 model. By using pre-trained SMG, the unaugmented pre-trained model perform better as we have fewer samples in the pre-trained SMG than target dataset iMiGUE. The moment-based augmented pre-trained SMG model also slightly underperforms with a difference of 1.15% than its unaugmented SMG model because of the low number of samples and different topologies could leads to model overfitting.

Even though the pre-trained model datasets sample sizes and skeleton topologies are different than the target model datasets, the overall cross-dataset evaluation performes better.

V. Conclusions

In this paper, we proposed an unsupervised approach to learn gestures representation more effectively using unlabeled skeleton data. The augmented sequence enables the model to learn more pattern invariant and discriminate features from unlabeled data. An efficient dense-graph convolutional network encoder along with a momentum-based efficient encoder and queue-based dictionary helps learn better gestures representation using a query, positive and negative keys. The learned gestures representations were fed to the linear classifier with their original labels for final micro-gestures recognition. The moment-based augmentation was used to improve the performance. We evaluated our proposed model using two micro-gestures datasets, SMG and iMiGUE and achieved the Top-1 accuracy of 47.9% and 37.5%, respectively. The cross-dataset experiments also exhibited the model's robustness. The improvement in accuracy of SMG dataset by 7.4% and iMiGUE dataset by 18.4% compared to their baselines is a substantial performance by unsupervised framework.

Acknowledgment

This work was supported by the Academy of Finland for Academy Professor project EmotionAI (grants 336116, 345122), project MiGA (grant 316765) and ICT 2023 project (grant 328115). The authors also wish to acknowledge Muzammil Behzad & Hanglin Mo for their assistance and CSC – IT Center for Science, Finland, for computational resources.

References

[1] L. Wang, D. Q. Huynh, and P. Koniusz, "A comparative review of recent kinect-based action recognition algorithms," *IEEE Transactions on Image Processing*, vol. 29, pp. 15–28, 2019.

[2] N. Zheng, J. Wen, R. Liu, L. Long, J. Dai, and Z. Gong, "Unsupervised representation learning with long-term dynamics for skeleton based action recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[3] 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.

[4] C. Si, W. Chen, W. Wang, L. Wang, and T. Tan, "An attention enhanced graph convolutional lstm network for skeleton-based action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1227–1236.

[5] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Two-stream adaptive graph convolutional networks for skeleton-based action recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 12 026–12 035.

[6] L. Devillers, L. Vidrascu, and L. Lamel, "Challenges in real-life emotion annotation and machine learning based detection," *Neural Networks*, vol. 18, no. 4, pp. 407–422, 2005.

[7] X. Liu, H. Shi, H. Chen, Z. Yu, X. Li, and G. Zhao, "imigue: An identity-free video dataset for micro-gesture understanding and emotion analysis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 10 631–10 642.

[8] H. Chen, X. Liu, X. Li, H. Shi, and G. Zhao, "Analyze spontaneous gestures for emotional stress state recognition: A micro-gesture dataset and analysis with deep learning," in *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*. IEEE, 2019, pp. 1–8.

[9] Z. Liu, H. Zhang, Z. Chen, Z. Wang, and W. Ouyang, "Disentangling and unifying graph convolutions for skeleton-based action recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 143–152.

[10] K. Su, X. Liu, and E. Shlizerman, "Predict & cluster: Unsupervised skeleton based action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9631–9640.

[11] L. Lin, S. Song, W. Yang, and J. Liu, "Ms2l: Multi-task self-supervised learning for skeleton based action recognition," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 2490–2498.

[12] J. Li, Y. Wong, Q. Zhao, and M. S. Kankanhalli, "Unsupervised learning of view-invariant action representations," *Advances in Neural Information Processing Systems*, vol. 31, pp. 1254–1264, 2018.

[13] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.

[14] R. Zhang, P. Isola, and A. A. Efros, "Split-brain autoencoders: Unsupervised learning by cross-channel prediction," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1058–1067.

[15] M. Gutmann and A. Hyvärinen, "Noise-contrastive estimation: A new estimation principle for unnormalized statistical models," in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 2010, pp. 297–304.

[16] Y. Tian, D. Krishnan, and P. Isola, "Contrastive multiview coding," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*. Springer, 2020, pp. 776–794.

[17] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9729–9738.

[18] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "Ntu rgb+ d: A large scale dataset for 3d human activity analysis," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1010–1019.

[19] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev *et al.*, "The kinetics human action video dataset," *arXiv preprint arXiv:1705.06950*, 2017.

[20] Y. Luo, J. Ye, R. B. Adams, J. Li, M. G. Newman, and J. Z. Wang, "Arbee: Towards automated recognition of bodily expression of emotion in the wild," *International journal of computer vision*, vol. 128, no. 1, pp. 1–25, 2020.

[21] 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.

[22] D. Yang, Y. Wang, A. Dantcheva, L. Garattoni, G. Francesca, and F. Bremond, "Unik: A unified framework for real-world skeleton-based action recognition," *arXiv preprint arXiv:2107.08580*, 2021.

[23] H. Rao, S. Xu, X. Hu, J. Cheng, and B. Hu, "Augmented skeleton based contrastive action learning with momentum lstm for unsupervised action recognition," *Information Sciences*, vol. 569, pp. 90–109, 2021.

[24] J.-R. Mor, Z. Alon, and R. Kimmel, "Momen^{e}t: Flavor the moments in learning to classify shapes," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019.