engineering proceedings

MDPI

Proceeding Paper

Semi-Supervised Adaptation for Skeletal Data Based Human Action Recognition [†]

Haitao Tian [1,*] and Pierre Payeur [2]

[1] University of Ottawa
[2] University of Ottawa; ppayeur@uottawa.ca
[*] Correspondence: htian026@uottawa.ca

[†] Presented at the 10th International Electronic Conference on Sensors and Applications (ECSA-10), 15–30 November 2023; Available online: https://ecsa-10.sciforum.net/.

Abstract: Recent research on human action recognition is largely facilitated by skeletal data, a compact representation composed of key joints of the human body. However, leveraging the capabilities of artificial intelligence on such sensory input imposes the collection and annotation of a large volume of skeleton data, which is extremely time-consuming. In this paper, a two-phase semi-supervised learning approach is proposed to surmount the high requirements on labeled skeletal data while training a capable human action recognition model adaptive to a target environment. In the first phase, an unsupervised learning model is trained under a contrastive learning fashion to extract high-level human action semantic representations from unlabeled source dataset. The resulting pre-trained model is then fine-tuned on a small number of properly labeled data of the target environment. Experimentation is conducted on large-scale human action recognition datasets to evaluate the effectiveness of the proposed method. Code is available at https://github.com/tht106/SSA.

Keywords: skeletal action recognition; semi-supervised learning; contrastive learning; domain adaptation

1. Introduction

As a notoriously data-driven learning technique, deep learning has demonstrated remarkable effectiveness in human action recognition by involving massive training on large-scale human activity datasets [1,2]. Recently, Graph Convolutional Networks (**GCN**s), tailored for skeleton data, which is efficiently extracted from 3D imaging sensors and that offers the merit of being robust to variations in the environment, have achieved state-of-the-art performance in human action recognition research [3,4].

Even though it is promising to realize a powerful human action recognition model via leveraging large-scale public datasets, the operation of the resulting model in practice could be challenging. Deploying the model into a target environment, where the distributions of the target data deviate partially from that of the source training data domain due to the particular imaging configuration adopted, refers to the data domain covariate issue in the deep learning community [5]. A common strategy to tackle the problem is to conduct extra fine-tuning rounds with full supervision of the data collected under the target imaging configuration, in order to eliminate the data discrepancy from the source domain to the target domain. However, the collection and annotation of a large volume of skeleton data for fine-tuning is extremely time-consuming. Meanwhile, the data collection in real environments could be highly restricted due to particular privacy considerations.

This work focuses on an important perspective related to practical deployment of a human action recognition model. It is related to the efficient adaptation of a GCN model from a public dataset domain to a target environment where only a limited number of data will be available for model refining. The semi-supervised adaptation strategy

Citation: Tian, H.; Payeur, P. Semi-Supervised Adaptation for Skeletal Data Based Human Action Recognition. *Eng. Proc.* **2023**, *56*, x. https://doi.org/10.3390/xxxxx

Academic Editor(s): Name

Published: 15 November 2023

Copyright: © 2023 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

effectively circumvents the overfitting issue in learning with small number of samples. In turn, it guarantees a target-environment aware action recognition model that is compatible with state-of-the-art performance. In a *first* stage, an unsupervised learning framework is proposed to discover high-level human action semantic patterns from plenty of unlabeled skeletal data samples of the source data domain. It is inspired by the recent advances in contrastive learning [6] which provides pivotal competency for learning domain invariant representations from unlabeled data. The unsupervised learning phase does not aim to realize a capable action recognition model but to develop an action pattern aware pretrained model for the next stage. In the *second* stage, the adaptation strategy refines the pretrained model on a small number of labeled skeletal samples to learn a target domain specific prediction model.

The contribution of the proposed work consists of two aspects. *First*, it investigates the underlying principles of the skeletal data distribution shift issue during the practical action recognition model deployment. *Second*, it investigates domain adaptation strategies to improve the action recognition models' generalizability and robustness by introducing a semi-supervised adaptation strategy which leads to significant reduction on data requirement in the target domain while achieving convincing performance.

2. Related Work

The research on human action recognition addresses a variety of downstream computer vision-based tasks, such as human activity analysis, anomaly action detection, and video surveillance in hazardous places. Recently, Graph Convolutional Networks (GCNs) demonstrated the capability of interpreting topological features from multi-dimensional skeleton sequences, thereby dominating the research on skeletal human action recognition [3,4,7]. Yet, the generalization of the action recognition models in real environments is still challenging due to the data distribution shift caused by the variations in sensory configuration, e.g., camera views, heights, orientations, locations, and variations in data collection. Although the current skeletal datasets are devoted to covering the skeletal variations during data collection (e.g., NTU RGB+D [1] involves data variations by configuring 3 camera angles, 16 differences in height and distance, as well as involving 40 actors into action performance), the expected robustness of the resulting action recognition model remains vulnerable and can be uprooted while facing domain shift in skeletal data [8].

Contrastive learning formulates unsupervised representations learning by contrasting positive pairs against negative pairs from a pre-defined dynamic dictionary [9,10]. In skeletal action recognition, the contrastive learning scheme regards each skeletal sequence as a unique class represented by an GCN encoder and exploits the skeletal invariances and similarities from the dynamic dictionary formulated by skeleton sequences without using data annotations [6]. Inspired by such advances in contrastive learning, this work is devoted to a semi-supervised adaptation scheme for human action recognition.

3. Method

This section defines the skeletal data distribution shift issue involved in the practical deployment of a human action recognition model. It then proposes a two-phase semi-supervised training framework, depicted in Figure 1, that relies on significant quantity of public data (unlabeled) for pretraining and then refines a target-adaptive model by using only a small number of data (labeled) of the target environment.

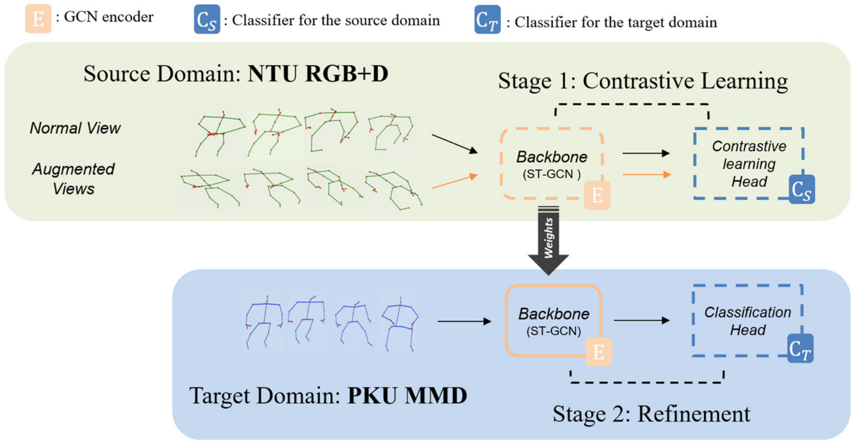

Figure 1. Framework of the proposed semi-supervised adaptation strategy. In the first stage, the training data (unlabeled) from the source domain is utilized to contrastive learning after data augmentation. The learned backbone is recycled in stage 2 for refining over the data samples (labeled) in the target domain. Network embedding is denoted with dotted lines if it is updated during training, and with solid lines otherwise.

3.1. Data Domain Shift in Skeletal Data

A human action recognition model considers the training dataset $\{X\}$ where $X \in \mathbb{R}^{T\times V\times 3}$ denotes a skeletal sequence composed of T frames in the shape of V human body joints, each one being defined in a calibrated camera reference frame with three-dimensional coordinates. $Y \in \mathbb{R}^{L}$ denotes the action label of X in a range of L categories. The goal of skeletal action recognition is to train a GCN model, composed of a graph convolution encoder **E** and a classification layer **C**, given the input (X_{train}, Y_{train}), where X_{train} is uniformly sampled from the dataset $\{X\}$. Normally, considering the test data sampled from $\{X\}$, i.e., the training and the test data are *i.i.d.* (independent identical distributions), the model achieves convincing evaluation performance on the test dataset. However, in practical engineering applications, the target deployment environment always presents misaligned data distributions given the fact that the imaging configuration can differ according to the environment, which leads to variations on data distribution such that X_{train} and X_T are not *i.i.d.*, where X_T defines the skeletal data sampled in the target environment $\{X_T, Y_T\}$. The data domain shift corrupts the model performance while the model was well-trained on $\{X_{train}\}$ but evaluated on $\{X_T\}$.

3.2. Semi-Supervised Learning

To effectively tackle the underlying issue of skeletal domain shift, **a two-stage strategy** is proposed which (i) utilizes sufficient public dataset samples to pretrain a GCN encoder **E** with contrastive learning, and then (ii) recycles the pretrained **E** in the second stage to refine a target-specific classification layer, $\mathbf{C_T}$, exclusively on the target domain samples. The learnt **E** and $\mathbf{C_T}$ reassemble the target domain adaptive action recognition model. The details are as follows.

- **In the first stage**, the Extremely Augmented Skeleton (EAS) scheme [6] is used to augment the training data with $\{X_{train}^{aug}\}$ to enrich the input space in both spatial and temporal dimensions via eight augmentation operations: *spatial shear, spatial flip, axis-wise rotate and mask, temporal flip, temporal crop, Gaussian noise and Gaussian blur*.

The input query-key pairs $(X_{train}, X_{train}^{aug})$ compose a dynamic skeletal dictionary upon which the skeleton contrastive learning framework learns underlying topological invariances and semantic similarities from the unlabeled source domain $\{X\}$. The training progress is driven by MoCov2 [11] with the InfoNCE loss:

$$L_{\mathrm{IN}} = -log \frac{\exp(Z_q \cdot Z_k)/\tau.}{\exp(Z_q \cdot Z_k)/\tau + \sum_{Z_n \in \mathcal{N}} \exp(Z_q \cdot Z_n)/\tau} \tag{1}$$

where Z_q denotes the feature representations of the query input X_{train} as $Z_q = \mathbf{C}_S(\mathbf{E}(X_{train}))$, where $\mathbf{C}_S$ denotes the multilayer perception (MLP) projection head for contrastive learning. Likewise, Z_k denotes the feature representation of the key input X_{train}^{aug} obtained by $\bar{\mathbf{C}}_S(\bar{\mathbf{E}}(X_{train}^{aug})$, where $\bar{\mathbf{C}}_S$ and $\bar{\mathbf{E}}$ are the mean models of $\mathbf{C}_S$ and $\mathbf{E}$, respectively. $\mathcal{N}$ represents negative sample representations to the query input. τ acts as the temperature parameter of the softmax operation in Equation (1).

- **In the second stage**, the pretrained model is refined on a small number of labeled skeleton data $\{X_T, Y_T\}$ of the target environment. It takes the well-learnt skeletal knowledge aware GCN encoder $\mathbf{E}$ from the first stage and fine-tunes the reassembled GCN model $\mathbf{C}_T(\mathbf{E}(\cdot))$ over the target domain. This stage is driven by a Cross-Entropy loss:

$$L_{\mathrm{CE}} = -Y_T \cdot \log[\, \mathbf{C}_T(\mathbf{E}(X_T))] \tag{2}$$

Note that the encoder $\mathbf{E}$ is free of gradients backpropagation in the second stage (as illustrated in Figure 1). As the encoder is pretrained, this fine-tuning process will quickly converge to the target domain data distribution and learn a capable action prediction layer.

4. Experiments

Comprehensive experimentation is conducted to evaluate the effectiveness of the proposed two-stage method in a cross-domain action recognition scenario.

4.1. Datasets and Implementations

This study employs two public datasets to simulate the skeletal distribution shift scenario. In particular, NTU RGB+D [1] is considered as the training dataset and PKU-MMD [2] as the target environment. The former is a popular large-scale skeleton form human action recognition dataset. It presents 56,880 samples covering 60 human daily actions recorded in indoor scenes with three cameras mounted in different locations to support variations in camera views. It is common to utilize such a large-scale dataset in the research community [3–5] for human action recognition. PKU-MMD is another popular public skeletal dataset presenting fewer data samples (20,000 instances over 52 actions), but it involves significant camera view variations in the samples. In this paper, it is employed to mimic practical environments where data collection configurations could be inconsistent. The samples related to 50 actions common to the two datasets are utilized for experimentation. In the first stage, the selected part from the NTU RGB+D dataset is used for unsupervised contrastive learning, while one tenth of the samples (namely 1884) of the PKU-MMD dataset is used for model refining in the second stage. Given the GCN architecture considered, ST-GCN [3] is adopted for the network backbone. The output dimension of $\mathbf{C}_S$ and $\mathbf{C}_T$ is set as 128 and 50, respectively. The parameter τ in Equation (1) is set as 0.07. For model training, it follows the same optimization details as utilized in [6].

4.2. Results

Experimental results are summarized in Table 1 that reports the Top-1 accuracy [1] of the resulting models while tested on the full test set of PKU-MMD. As a comparative, a source-only model is trained with exclusive full supervision using the NTU RGB+D training dataset with annotations. The results in the first row of Table 1 illustrate that the resulting model yields convincing results on the NTU RGB+D benchmark, while that source only model fails to effectively transfer to the target domain, demonstrating the negative impacts of domain shift on model performance.

Table 1. Comparative performance of the proposed method against full supervised training.

Model	On Benchmark (NTU RGB+D)	On Target Domain (PKU-MMD)
Source only	**69.07%**	57.32%
Adaptation	34.41%	**75.74%**

Next, the model is adapted by utilizing the proposed semi-supervised learning scheme. Results in the second row of Table 1 reflect the significant improvement achieved with the proposed method on the target domain, achieving a gain of 18.42% (from 57.32% to 75.74%). We conjecture that, as the data annotations of the source domain are not utilized in contrastive learning, the learnt encoder **E** is action semantic aware but not overfitting to the source domain, which finally results in a capable human action model for PKU-MMD. However, a severe performance deterioration is also observed on the benchmark NTU RGB+D evaluation (from 69.07% to 34.41%). It demonstrates that the encoder **E** does not form a competent action recognition stage on the source data domain but rather a reliable intermediate action semantic aware encoding which is efficiently generated by unsupervised contrastive learning.

With the goal to identify the principle that drives performance increase with respect to the different amounts of target domain samples involved for model refining, Table 2 reports on the results of an experiment where the encoder **E** learnt in the first stage remains fixed but the classification layer $\mathbf{C}_T$ is refined upon different ratios (varying from 5% to 100%) of samples selected from the target domain PKU-MMD. Experimental results demonstrate that the model performance tends to increase monotonically along with the ratio of the refining data samples. Interestingly, even using a very small number (e.g., 5%) of data samples from the target domain, the model still achieves convincing performance compared to the best model (85.06% in Table 2) when using 100% of the PKU-MMD dataset. In conclusion, this study suggests good trade-off conditions between data usage and performance gains while utilizing the proposed two-stage semi-supervised method.

Table 2. Performance gains related to different proportions of the target data samples for $\mathbf{C}_T$ refinement training.

Percentage of data use	5%	20%	30%	50%	70%	100%
Accuracy	71.60%	77.33%	81.03%	82.25%	83.28%	85.06%

4.3. T-SNE Action Clusters Visualization

For better understanding of the effectiveness of the proposed method, closer examination of the embedding features of the two domains is presented using t-SNE [12]. It is expected that a capable classification model can interpret separable and dense action clusters on the feature space. Specifically, Figure 2 visualizes the action clusters of the two respective domains on the last convolutional layer before the classification head as interpreted by the two ST-GCN models (“Source only” on the upper row and “Adaptation” at the bottom). The “Source only” model presents well separated action clusters when tested on the source domain (left column), which reflects that the model is able to interpret feature representations from the source domain on which the classification head easily determines action-wise classification boundaries. However, under the impacts of domain shift, the source-only model presents less separable action clusters when tested on the target domain (right column). Such a discrepancy on action cluster interpretation leads to the performance difference across the two domains (69.07% vs. 57.32% in Table 1). Conversely, after applying the proposed semi-supervised learning scheme as a refinement stage, the model (bottom row) demonstrates effective adaptation to the target domain whose action clusters are improved in terms of separability. Improvement is observable on both tests considering the source domain (left column) and the target domain (right column).

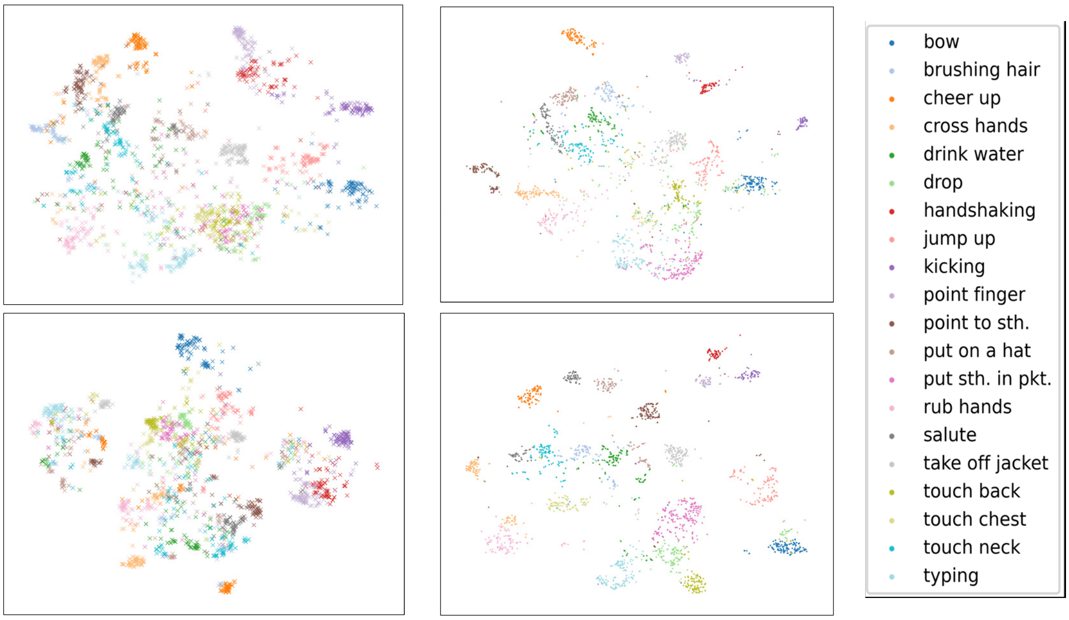

Figure 2. T-SNE visualization on action clusters on the embedding space of ST-GCN (**upper row**: "Source only"; **bottom row**: "Adaptation"). Clusters are distinguished by colors where twenty actions are randomly selected among fifty actions for clarity. Left column represents action clusters of the source domain (NTU RGB+D) and right column shows clusters of the target domain (PKU-MMD), respectively.

4.4. Semi-Supervised Learning vs. Supervised Learning

This subsection presents two experiments to evaluate whether utilizing fully supervised learning can reach the same effectiveness as the proposed semi-supervised adaptation method. First, a model trained with full supervision from NTU RGB+D is then fine-tuned also with full supervision on 10% data samples from PKU-MMD. Second, a separate model is trained with full supervision over the combined data samples from NTU RGB+D and PKU-MMD (10%). Both trained models are tested on the test dataset of PKU-MMD. Experimental results in Table 3 demonstrate that either fully supervised learning method presents inferior performance compared to the proposed semi-supervised method. Especially, the fully supervised transfer learning (pretrain on NTU RGB+D and fine-tuned on PKU-MMD) only achieves 45.97% on the target domain, representing a 29.77% gap in accuracy compared to the proposed method.

Table 3. Performance of Fully supervised learning vs. Semi-supervised learning.

	Full Supervision		**Semi-Supervision**
	NTU RGB+D & 10% PKU-MMD (Fine-Tuning)	**NTU RGB+D & 10% PKU-MMD (Combined)**	**NTU RGB+D & 10% PKU-MMD (Fine-Tuning)**
Accuracy	45.97%	62.61%	75.74%

5. Conclusions

This work proposes a simple but efficient method to deploy a skeleton data based human action recognition model to a target environment while requiring only a small amount of labeled data from the latter. The proposed semi-supervised learning strategy utilizes contrastive learning to pretrain a model that learns key skeletal representations from an unlabeled dataset, then fine-tunes the pretrained model on a small number of labeled data samples in the target domain. Experiments are conducted to demonstrate the effectiveness of the proposed strategy. It suggests that the semi-supervised learning method achieves convincing results compared to fully supervised learning that requires voluminous labeled data from both the source and target domains. The research also experimentally characterizes a tradeoff between data usage and model performance, providing reference to develop and deploy future applications.

Author Contributions: Conceptualization, H.T. and P.P.; methodology, software, validation, formal analysis, H.T.; investigation, H.T. and P.P.; resources, P.P.; writing—original draft preparation, H.T.; writing—review and editing, P.P.; supervision, project administration, funding acquisition, P.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported in part by MITACS Accelerate and NSERC Discovery grants.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Shahroudy, A.; Jun, L.; Tian-Tsong, N.; Gang, W. NTU RGB+D: A large scale dataset for 3d human activity analysis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1010–1019.
2. Liu, C.; Hu, Y.; Li, Y.; Song, S.; Liu, J. PKU-MMD: A large scale benchmark for skeleton-based human action understanding. In Proceedings of the Workshop on Visual Analysis in Smart and Connected Communities, Mountain View, CA, USA, 23 October 2017; pp. 1–8.
3. Yan, S.; Xiong, Y.; Lin, D. Spatial temporal graph convolutional networks for skeleton-based action recognition. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; Volume 32.
4. Chen, Y.; Zhang, Z.; Yuan, C.; Li, B.; Deng, Y.; Hu, W. Channel-wise topology refinement graph convolution for skeleton-based action recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 13359–13368..
5. Ben-David, S.; Blitzer, J.; Crammer, K.; Kulesza, A.; Pereira, F.; Vaughan, J.W. A theory of learning from different domains. *Mach. Learn.* **2010**, *79*, 151–175.
6. Guo, T.; Liu, H.; Chen, Z.; Liu, M.; Wang, T.; Ding, R. Contrastive learning from extremely augmented skeleton sequences for self-supervised action recognition. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual, 22 February–1 March 2022; Volume 36, pp. 762–770.
7. Chi, H.G.; Ha, M.H.; Chi, S.; Lee, S.W.; Huang, Q.; Ramani, K. InfoGCN: Representation learning for human skeleton-based action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 20186–20196.
8. Choi, J.; Sharma, G.; Chandraker, M.; Huang, J.-B. Unsupervised and semi-supervised domain adaptation for action recognition from drones. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Snowmass Village, CO, USA, 1–5 March 2020; pp. 1717–1726.
9. Chen, T.; Kornblith, S.; Norouzi, M.; Hinton, G. A simple framework for contrastive learning of visual representations. In Proceedings of the International Conference on Machine Learning, Virtual, 13–18 July 2020; pp. 1597–1607.
10. He, K.; Fan, H.; Wu, Y.; Xie, S.; Girshick, R. Momentum contrast for unsupervised visual representation learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 9729–9738.
11. Chen, X.; Fan, H.; Girshick, R.; He, K. Improved baselines with momentum contrastive learning. *arXiv* **2020**, arXiv:2003.04297.
12. Van der Maaten, L.; Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.