Contrastive Self-Supervised Learning for Skeleton Action Recognition

Xuehao Gao [1]
School of Software Engineering
Xi'an Jiaotong University
gaoxuehao.xjtu@gmail.com

Yang Yang [1*]
School of Automation Science and Engineering
Xi'an Jiaotong University
yyang@mail.xjtu.edu.cn

Shaoyi Du [2†]
College of Artificial Intelligence
Xi'an Jiaotong University
dushaoyi@xjtu.edu.cn

Abstract

Learning discriminative features plays a significant role in action recognition. Many attempts have been made to train deep neural networks by their labeled data. However, in previous networks, the view or distance variations can cause the intra-class differences even larger than inter-class differences. In this work, we propose a new contrastive self-supervised learning method for action recognition of unlabeled skeletal videos. Through contrastive representation learning by adequate compositions of viewpoints and distances, the self-supervised net selects discriminative features which have invariance motion semantics for action recognition. We hope this attempt can be helpful for the unsupervised learning study of skeleton-based action recognition.

1 Introduction

Action understanding is a fundamental study in the computer vision field. Compared with RGB images, body joint time-series (skeletons) are effective descriptors of actions, which are robust against the background and lighting changes. Since each joint is easily identified by a 2D or 3D position vector, skeleton sequence becomes a high-level and abstract representation of an action. Although recent methods have achieved remarkable progress with the development of deep neural networks, most methods rely on strong supervision on action labels. Large-scale data annotation is laborious and expensive, even impractical for complex data such as videos. Furthermore, annotation is a challenging problem by itself, since some action classes are ambiguous, which are up to the interpretation of each annotator for a given sequence. Thus, unsupervised methods that do not use labeled data are necessary for skeleton action recognition.

Learning effective motion representations without human supervision is a long-standing and challenging problem. Existing approaches [8] are mostly based on the encoder-decoder architecture. Specifically, given a skeleton action sequence as the encoder input, the decoder predicts the encoder's input sequence. But this idea needs to re-generate each frame in a sequence, which is computationally expensive and often ignores the semantic association between the encoder input and the decoder output. In that case, the generality of the learned representation is limited. Furthermore, the auto-encoder will not be able to achieve efficient performances without particular training strategies [8]. Inspired

[*] X.Gao and Y.Yang contributed equally to this work.
[†] Corresponding author.

Pre-registration workshop NeurIPS (2020), Vancouver, Canada.

by the recent successful discriminative approaches in the latent space [1, 2, 7], here we try to obtain effective feature representations of skeleton sequences under a simple contrast learning framework.

Previous approaches [3, 6] have proven that one of the key challenges in action recognition lies in the large variety of action representation when motions are captured from different viewpoints. But, in reality, the same action can be easily recognized by two observers when they stand at different viewpoints or distances. This observation inspires us to obtain effective feature representations under a contrastive self-supervised learning framework. Accordingly, we define contrastive prediction tasks by the composition of multiple data augmentation operations (*i.e.* view variations). And then we maximize the agreement between different augmented views of the same data example via a contrastive loss in the latent space. Through contrastive representation learning by adequate compositions of viewpoints and distances, the self-supervised net selects discriminative features which have invariant motion semantics for action recognition.

2 Related Work

Numerous approaches have been introduced particularly for human action recognition. Most of them are supervised where an annotated set of actions and labels should be provided for training. In an unsupervised setup, the problem of action recognition is much more challenging and only a few methods have been proposed. In this section, we give a brief review on the unsupervised method and the skeleton action recognition by variant of viewpoint and distance.

2.1 Unsupervised Learning in Skeleton Action Recognition

The unsupervised setup has advantages for action recognition since it does not require the labeling of sequences. When additional action appears, the unsupervised networks do not need re-training. This kind of method typically aims to obtain an effective feature representation by predicting or regenerating future frames of input sequences. Srivastava et al. [4] proposed a recurrent-based sequence (Seq2Seq) model as an autoencoder to learn the representation of a video. Zheng et al. [5] proposed a GAN encoder-decoder(LongTGAN). The decoder attempts to re-generate the input sequence and the discriminator is used to discriminate whether the re-generation is accurate. Su et al. [8] make the decoder and the encoder self-organize their hidden states into a feature space which clusters similar movements into the same cluster and distinct movements into distant clusters. The feature representation used for action recognition is taken from the final state of the encoder's hidden representation. However, these typical encoder-decoder-based unsupervised methods are computationally expensive and rely on particular training strategies. We explore to learn robust and efficient features representation based on a simpler yet effective idea in the contrastive learning framework.

2.2 Variant of Viewpoints and Distance in Skeleton Action Recognition

For humans, it will not affect us to recognize the action class from arbitrary observing viewpoints and distances. Human actions can be captured from arbitrary camera setups, while it is a challenging problem for recognition algorithms. Researchers have paid much attention to this issue and proposed various view-invariant approaches. Zhang et al. [3] proposed a view-adaptive idea to leverage content-dependent and view-adaptation models to automatically learn and determine the suitable viewpoint. And for each sequence, the skeletons are transformed into representations under those views. This improvement has proved that viewpoints are crucial for the model to learn features representation in skeleton-based action recognition. Furthermore, the distance of a subject to the camera may influence the scale of the skeletons. In order to learn the robust and effective features representation in an unannotated skeleton sequence, we design a simple contrastive learning method based on the idea of semantic invariance under the variant of distance and viewpoint.

3 Methodology

We learn our feature representations by maximizing agreement between different viewpoints and distance setups of the same skeleton sequence via the contrastive loss in the latent space. As

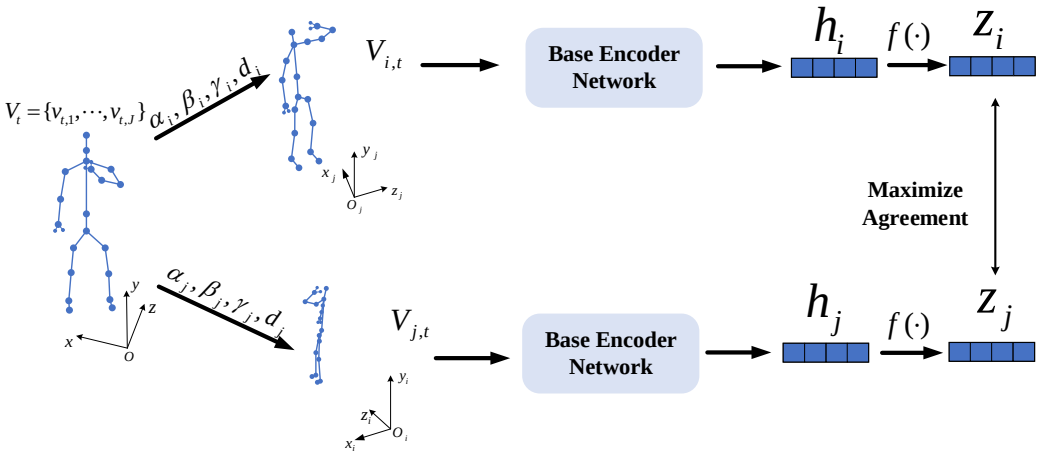

Figure 1: The framework for contrastive learning of skeleton sequence using the semantic invariance under the variant of distance and viewpoint. Two separate transformations of viewpoint and distance are applied to a given skeleton sequence to obtain a correlated pair. A base encoder network and a projection function $\boldsymbol{f}(\cdot)$ are trained to maximize agreement by using a contrastive loss.

illustrated in Figure 1, we just using the base encoder network (*e.g.* $ResNet$ [11] and GCN [12]) and representation vector $\boldsymbol{h}$ for following action recognition task.

3.1 View Augmentation for Contrastive Representation Learning

The raw 3D skeleton is recorded corresponding to the camera coordinate system (global coordinate system), with the origin located at the position of the camera sensor. Note that the input skeleton $\boldsymbol{V_t}$ to our system as in Figure 1 is the skeleton representation under this initial camera coordinate system. To be insensitive to the initial position of action, for each sequence, we translate the global coordinate system to the body center of the first frame as our new global coordinate system. One can choose to observe the action from stochastic positions. Thanks to the availability of the 3D skeletons captured from a fixed view, it is possible to set up a movable virtual camera and observe the action from new observation viewpoints. The given skeleton can be transformed into a representation under the movable virtual camera coordinate system, which is also referred to as the observation coordinate system.

Given a skeleton sequence $\boldsymbol{S}$ with $\boldsymbol{T}$ frames, under the global coordinate system, we denote the set of joints in the $\boldsymbol{t}$-th frame as $\boldsymbol{V_t} = \{\boldsymbol{v_{t,1}}, \cdots, \boldsymbol{v_{t,J}}\}$. For the $\boldsymbol{t}$-th frame, we assume the movable virtual camera is placed at a stochastic viewpoint and distance, with the corresponding observation coordinate system obtained from a translation by $\boldsymbol{d}_t \in \mathbb{R}^3$, and a rotation of $\boldsymbol{\alpha_t}, \boldsymbol{\beta_t}, \boldsymbol{\gamma_t}$ radians anticlockwise around the $\boldsymbol{x}$-axis, $\boldsymbol{y}$-axis, and $\boldsymbol{z}$-axis, respectively, of the global coordinate system. Therefore, the representation of the $\boldsymbol{j}$-th skeleton joint $\boldsymbol{v'_{t,j}} = \left[x'_{t,j}, y'_{t,j}, z'_{t,j}\right]^{\mathrm{T}}$ of the $\boldsymbol{t}$-th frame under this observation coordinate system $\boldsymbol{O}$ is

$$\boldsymbol{v'_{t,j}} = \left[x'_{t,j}, y'_{t,j}, z'_{t,j}\right]^{\mathrm{T}} = \mathbf{R}_t \times \left(\boldsymbol{v_{t,j}} - \boldsymbol{d}_t\right). \tag{1}$$

$\mathbf{R}_t$ can be represented as

$$\mathbf{R}_t = \mathbf{R}^x_{t,\alpha} \times \mathbf{R}^y_{t,\beta} \times \mathbf{R}^z_{t,\gamma}, \tag{2}$$

where $\mathbf{R}^y_{t,\beta}$ denotes the coordinate transformation for rotating the original coordinate system around the $\boldsymbol{y}$-axis by $\boldsymbol{\beta_t}$ radians anticlockwise, which is defined as

$$\mathbf{R}^y_{t,\beta} = \begin{bmatrix} \cos\left(\beta_t\right) & \sin\left(\beta_t\right) & 0 \\ -\sin\left(\beta_t\right) & \cos\left(\beta_t\right) & 0 \\ 0 & 0 & 1 \end{bmatrix}. \tag{3}$$

Similarly, $\mathbf{R}^x_{t,\alpha}$ and $\mathbf{R}^z_{t,\gamma}$ denote the coordinate transforms for rotating the original coordinate system around the $\boldsymbol{x}$-axis by $\boldsymbol{\alpha_t}$ radians, and around the $\boldsymbol{z}$-axis by $\boldsymbol{\gamma_t}$ radians anticlockwise, respectively. Note that all the skeleton joints in the $\boldsymbol{t}$-th frame share the same transform parameters, *i.e.*,

$\boldsymbol{\alpha_t}, \boldsymbol{\beta_t}, \boldsymbol{\gamma_t}, \boldsymbol{d_t}$, considering that the changing of viewpoints and distance is a rigid transformation. Given these transformation parameters, the skeleton representation $\boldsymbol{V_t} = \{\boldsymbol{v_{t,1}}, \cdots, \boldsymbol{v_{t,J}}\}$ under the new observation coordinate can be obtained from Eq.1. Besides, the viewpoint and distance can be varied for different frames or on a sequence level. The key problem becomes how to obtain an effective feature representations in skeleton sequence under the contrast learning framework by adequate compositions of viewpoints and distances.

3.2 The Contrastive Learning Framework

Firstly, given a skeleton sequence $\boldsymbol{V_t}$ that is captured at an initial position and orientation, we adopt two stochastic transforms setups on it to provide a positive pair, denoted $\boldsymbol{V_{i,t}}$ and $\boldsymbol{V_{j,t}}$. Secondly, a base encoder network that extracts representation vectors $\boldsymbol{h}$ from transformed skeleton sequences. Similar to other typical contrastive learning methods, this simple framework design conveniently decouples the predictive task from other components such as the base encoder network architecture, and this base encoder network is shared by both branches of our framework. The base encoder network serves as the skeleton-based motion features extractor. Here, we can adopt the commonly used GCN [12]or $ResNet$ [11]scheme to extract spatial and temporal features. Finally, projection function $f(\cdot)$ will map representations to space where contrastive loss is applied. We opt for simplicity and use 2-layer MLP as a projection function. The base encoder network and projection function are trained to maximize the agreement by using contrastive loss. After that, we just use the pretrained base encoder network and the representation vector $\boldsymbol{h}$ for downstream tasks.

We randomly sample a minibatch of N examples and define the contrastive prediction task on pairs of view-augmented examples derived from the minibatch, resulting in $2N$ data points. For a given positive pair, we treat other $2(N-1)$ as negative samples. The contrastive loss function is used to maximize the agreement between positive pair, meanwhile minimizing the agreement between negative pairs. Here, we adopt a common formulation of loss function used in the previous contrastive learning work. Then the loss function of a positive pair of examples (i, j) is defined as

$$\ell_{i,j} = -\log \frac{\exp\left(\operatorname{sim}\left(\boldsymbol{z_i}, \boldsymbol{z_j}\right)/\tau\right)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp\left(\operatorname{sim}\left(\boldsymbol{z_i}, \boldsymbol{z_k}\right)/\tau\right)}, \tag{4}$$

where $\operatorname{sim}(\boldsymbol{z_i}, \boldsymbol{z_j}) = \boldsymbol{z_i}^\top \boldsymbol{z_j} / \|\boldsymbol{z_i}\| \|\boldsymbol{z_j}\|$ denotes the dot product between ℓ_2 normalized $\boldsymbol{z_i}$ and $\boldsymbol{z_j}$ (*i.e.* cosine similarity). $\mathbb{1}_{[k \neq i]}$ is an indicator function evaluating to 1 iff $k \neq i$ and τ denotes a temperature parameter. After that the final loss function is

$$\mathcal{L} = \frac{1}{2N} \sum_{k=1}^{N} [\ell(2k-1, 2k) + \ell(2k, 2k-1)]. \tag{5}$$

As illustrated in Eq.5, the final loss is calculated across all positive pairs, both (i, j) and (j, i).

4 Experimental Protocol

Here, we lay out the protocol for our empirical studies, which aims to understand different design choices in our framework and improve representation quality by combining our findings.

4.1 Dataset and Metrics

For learning feature representations without labels by the base encoder networks, we conduct unsupervised pretraining on NTU-RGB+D 60 and NTU-RGB+D 120. To evaluate the learned representations, we follow the widely used linear evaluation protocol, where a linear classifier is trained on the representation vector obtained from the frozen base neural network. The test accuracy is used as a proxy for the representation quality.

4.1.1 NTU-RGB+D 60 Dataset

The NTU RGB+D 60 [9] is a large-scale action recognition dataset collected by three Kinect cameras simultaneously. It contains 56,578 skeleton sequences and 60 action classes performed by 40 different subjects. Each human skeleton is represented by 25 joints with 3D coordinates. The authors

recommend evaluating the model performance under two settings: (1) Cross-Subject(X-Sub), where half of 40 subjects are used for training and the rest for testing. (2) Cross-View(X-view), where the sequences captured by two of three cameras are used for training, and those captured by the last cameras are used for testing.

4.1.2 NTU-RGB+D 120 Dataset

This dataset [10] is an extension of NTU-RGB+D 60 and currently the largest dataset for 3D action recognition. It contains 114,480 skeleton sequences and 120 action classes performed by 106 different subjects. It provides two types of evaluating setting: (1) Cross-Subject(X-Sub), where half of 106 subjects are used for training and the rest for testing. (2) Cross-Setup(X-Set), where the sequences captured by half of cameras are used for training and the rest for testing.

4.2 Default Setting

Benefitting from the experiences of several previous contrast learning methods, we realize that contrastive learning benefits from the larger batch size and more training steps compared to supervised learning. Thus, we will conduct training experiments with slightly larger batch sizes and steps purposefully. However, training with a large batch size may be unstable when using standard SGD/Momentum with linear learning rate scaling. To stabilize the training, we use the LARS optimizer for all batch sizes. Besides, in distributed training with data parallelism, the mean and variance of Batch Normalization (BN) are typically aggregated locally per device. In contrastive learning, as positive pairs are computed in the same device, the model can exploit the local information leakage to improve the prediction accuracy without improving representations. We address this issue by aggregating BN mean and variance overall devices during the training.

4.3 Variant of Viewpoints and Distance for Contrastive Representation Learning

Previous work has realized that viewpoint is crucial for features representation in supervised methods. Furthermore, the distance of a subject to the camera does influence the scale of the skeletons. To exploit this semantic invariance under the variants of distance and viewpoint, we conduct a contrastive learning framework for learning robust feature representations in an unannotated skeleton sequence. In the following experiments, we will adopt variants of viewpoints and distance to define the contrastive prediction task in a systematic way. We consider that frame-wise transformation is inferior to the sequence level transformation because the former loses more information, *e.g.* the motion across frames. Thus, we will conduct transformations on the sequence level in following experiments. To comprehensively study the impact of viewpoint and distance, we define positions of raw 3D skeletons are initial state, and provide several viewpoints and distance transformations, respectively. For each viewpoint setup, $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}$ are randomly selected from the degree set (*i.e.* $0°, 60°, 120°, 180°, 240°, 300°$). For each distance setup, $\boldsymbol{d}$ is a random times the unit distance where the alternative multiplier is $0 \sim 5\times$. To understand the effects of individual transformation of the skeleton and the importance of transformations composition, the experiment will be repeated multiple times:

- on stochastic viewpoints and distance setups with alternative $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{d}$ options.
- on stochastic viewpoints setups with random $\boldsymbol{\alpha}, \boldsymbol{\beta}$ pair and a fixed $\boldsymbol{\gamma}$.
- on stochastic viewpoints setups with random $\boldsymbol{\alpha}, \boldsymbol{\gamma}$ pair and a fixed $\boldsymbol{\beta}$.
- on stochastic viewpoints setups with random $\boldsymbol{\beta}, \boldsymbol{\gamma}$ pair and a fixed $\boldsymbol{\alpha}$.
- on stochastic distance setups with a fixed viewpoints setup.
- with base encoder architectures of different complexity (*i.e.* models with varied depth and width).

Acknowledgments and Disclosure of Funding

This work was supported by the National Key Research and Development Program of China under Grant No. 2018AAA0102500 and the National Natural Science Foundation of China under Grant No. 61971343. We would like to thank Maosen Li for his feedback on the draft.

References

[1] Bachman, P., Hjelm, R. D., & Buchwalter, W. (2019). Learning representations by maximizing mutual information across views. In Advances in Neural Information Processing Systems (pp. 15535-15545).

[2] Dosovitskiy, A., Springenberg, J. T., Riedmiller, M., & Brox, T. (2014). Discriminative unsupervised feature learning with convolutional neural networks. In Advances in neural information processing systems (pp. 766-774).

[3] Zhang, P., Lan, C., Xing, J., Zeng, W., Xue, J., & Zheng, N. (2019). View adaptive neural networks for high performance skeleton-based human action recognition. IEEE transactions on pattern analysis and machine intelligence, 41(8), 1963-1978.

[4] Srivastava, N., Mansimov, E., & Salakhudinov, R. (2015, June). Unsupervised learning of video representations using lstms. In International conference on machine learning (pp. 843-852).

[5] Zheng, N., Wen, J., Liu, R., Long, L., Dai, J., & Gong, Z. (2018, April). Unsupervised representation learning with long-term dynamics for skeleton based action recognition. In Thirty-Second AAAI conference on artificial intelligence.

[6] Li, J., Wong, Y., Zhao, Q., & Kankanhalli, M. S. (2018). Unsupervised learning of view-invariant action representations. In Advances in Neural Information Processing Systems (pp. 1254-1264).

[7] Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. arXiv preprint arXiv:2002.05709.

[8] Su, K., Liu, X., & Shlizerman, E. (2020). Predict & cluster: Unsupervised skeleton based action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 9631-9640).

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

[10] Liu, J., Shahroudy, A., Perez, M. L., Wang, G., Duan, L. Y., & Chichung, A. K. (2019). Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding. IEEE transactions on pattern analysis and machine intelligence.

[11] He, K. , Zhang, X. , Ren, S. , & Sun, J. . (2016). Deep Residual Learning for Image Recognition. IEEE Conference on Computer Vision & Pattern Recognition. IEEE Computer Society.

[12] Yan, S. , Xiong, Y. , & Lin, D. . (2018). Spatial temporal graph convolutional networks for skeleton-based action recognition.