Pose-based Action Recognition: Robust Models with Contrastive Loss

Student: Aleksandra Novikova
Supervisor: Mohamed Ossama Ahmed Abdelfattah
Semester project in VITA Lab, EPFL

I. Introduction

Pose-based action recognition is a cutting-edge task in computer vision. There are numerous different approaches to this task, and state-of-the-art solutions show great results [1], [2], [3]. However, these methods have several limitations, some of which we will attempt to explore and improve in this project.

Most methods, including state-of-the-art ones, work with heatmaps for human pose. This approach requires greater computational resources compared to predicting actions solely based on keypoints (pure skeleton). In this project, we will investigate the performance of models for different input data: heatmaps, skeletons, as well as grayscale images.

Another limitation of current models for this task is their sensitivity to noise, which can be a significant problem for performance on real-world data. To address this limitation and enhance the robustness of pose-based action recognition models, we present a novel approach leveraging contrastive loss. Contrastive learning has gained significant attention in the field of computer vision for its ability to learn powerful representations in a semi-supervised manner. We utilize this idea to incorporate noisy data (analogous to unlabeled data in semi-supervised learning). By exploiting the inherent structure and relationships within the data, contrastive loss allows the model to discern relevant features while ignoring irrelevant features, thereby improving generalization and robustness.

Our project focuses on redesigning existing pose-based action recognition models to make them more resilient to noise. We are exploring several state-of-the-art models combined with contrastive learning and noisy data on various input data formats. To assess the effectiveness of the resulting models, we conducted a series of experiments on the NTU RGB+D dataset, one of the most popular and extensive datasets in this field. Additionally, we compare the performance of the obtained models across different levels of input data noise.

The source code can be found on GitHub:
Repository

II. Related Work

A. Action Recognition

In recent years, the field of action recognition has been actively developing, and new advanced solutions for this task are emerging. Several notable works have focused on the use of neural networks to extract features from video frames. In one of the works [4], a two-stream architecture based on CNN was introduced, where spatial and temporal features are separately extracted. Later, in a number of works [5], [6], this idea was extended by transforming 2D convolutions into 3D convolutions to simultaneously capture spatial-temporal features.

More recently, in 2021, a new model called PoseConv3D [1] was presented, which achieves outstanding results on various datasets and is considered the state-of-the-art model for the NTU RGB+D dataset [7]. This model also utilizes the idea of 3D convolutions to capture spatial-temporal pose changes. It takes 3D heatmaps based on the skeleton as input.

Another modern solution is the MotionBERT model [3]. The authors of this model have provided a universal motion encoder that recovers 3D motion based on 2D skeletons. The encoder also employs a two-stream architecture and extracts geometric, kinematic, and physical motion features. Thus, after feature extraction, it can be fine-tuned for various computer vision tasks, including action recognition.

A brief comparison of these two models is presented in Table I.

TABLE I: Comparison of models PoseConv3D and MotionBERT, accuracy is presented for the NTU RGB+D dataset

Model	Input	Top-1 acc	Robustness
PoseConv3D	Heatmaps	93.1	No
MotionBERT	Skeletons	93.0	No

B. Contrastive Learning

Contrastive learning is widely used in the field of computer vision, including the task of action recognition. It allows models to be trained on large amounts of unlabeled data. Contrastive learning aims to map similar instances closer to each other and push dissimilar instances apart in a latent space, thereby facilitating the discrimination between different actions. In the work [2], the TCL model is presented, which utilizes a two-stream model for unlabeled data: one stream for high-frequency frames and another for low-frequency frames. Using contrastive loss, they maximize the similarities between representations of the same videos but with different frame frequencies and minimize the similarity between different videos.

III. Methodology

For our research, we used the PoseConv3D and MotionBERT models as backbones and the architecture of the TCL model for incorporating contrastive loss.

A. Different inputs

Initially, we investigated the differences between various input data types. Since PoseConv3D takes 3D heatmaps as input, while the MotionBERT model utilizes 2D skeletons, we decided to modify the PoseConv3D model to also work with pure skeletons. We modified the dataset class to enable the model to handle any of the three types of data: pure skeletons sized 17×2 (17 keypoints with 2 coordinates for each point), 3D heatmaps based on skeletons sized $17 \times 56 \times 56$ (17 heatmaps sized 56×56 for each keypoint), and grayscale images sized 56×56 (constructed by merging the heatmaps together).

B. Model with contrastive loss

For the robustness analysis of our models, we implemented a model similar to the architecture of the TCL model for semi-supervised learning. However, instead of using unlabeled data with different frame frequencies, we used the original skeletons and skeletons with added noise: we add zero-mean gaussian noise to coordinates of the skeleton with $std = \alpha$. For the noise, we use the parameter α, which represents the percentage of the original video size by which we want to shift the initial keypoint. Additionally, this parameter is multiplied by the normalization parameter, $total\_difference$, which is a value between 0 and 1 indicating the degree of change in action over time. For motionless actions (e.g., when only the hands are moving), this value is closer to 0, whereas for significant actions (e.g., when the entire body is involved), this value is closer to 1.

The architecture of the resulting model can be seen in Figure 1. There are a total of 3 streams: 1 stream for supervised learning, where the original non-noisy labeled data is used. For this stream, the standard cross-entropy loss is computed.

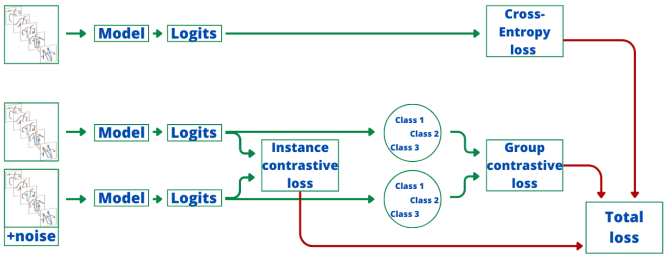

Fig. 1: Model architecture with contrastive loss

The other 2 streams are for contrastive learning. The first stream also uses clean non-noisy data, while the second stream uses the input data with added noise.

Contrastive loss is computed in two stages: instance contrastive loss and group contrastive loss.

The instance contrastive loss is defined as follows:

$$L_{inst}(S_o^i, S_n^i) = -\log \frac{h(S_o^i, S_n^i)}{\sum_k h(S_o^i, S_n^k) + \sum_{k \neq i} h(S_o^i, S_o^k)}$$

where h - is the exponential of cosine similarity measure, S_o^i, S_n^i - representation of the i-th skeleton for non-noisy and noisy data, respectively.

Thus, it brings the original skeletons and the noisy skeletons for the same video closer together, while pushing apart the skeletons from different videos.

The group contrastive loss is defined as follows:

$$L_{gr}(G_o^i, G_n^i) = -\log \frac{h(G_o^i, G_n^i)}{\sum_k h(G_o^i, G_n^k) + \sum_{k \neq i} h(G_o^i, G_o^k)}$$

where h - is the exponential of cosine similarity measure, G_o^i, G_n^i - average representations of the i-th class(group).

This loss brings the non-noisy skeletons closer to the noisy skeletons of the same class, and pushes apart skeletons from different classes.

The final loss is computed as the sum of the three losses with coefficients:

$$L_{total} = L_{cross-entropy} + \gamma_1 * L_{inst} + \gamma_2 * L_{gr}$$

IV. Experiments

For all experiments, we used the Cross-subject (X-Sub) data from the NTU RGB+D dataset. This dataset contains 60 different actions, about 40K train videos and about 16K validation videos. In each experiment, for each video, we uniformly sampled 24 frames from it. We trained each model for 150 epochs with dropout $p = 0.5$ and with learning rate 0.001. The remaining parameters varied depending on the experiment and will be described later.

A. Supervised learning

For experiments with different input data, we used the PoseConv3D model. As described earlier, the input data consisted of three types: pure skeletons, 3D heatmaps, and grayscale images. The experiment results are described in Table II.

TABLE II: PoseConv3D model results on NTU RGB+D dataset for different input data types

Input type	Top-1 acc	Top-5 acc
Heatmaps	90.1	99.1
Grayscale	88.7	98.8
Skeleton	87.5	98.7

The results for all types of input data are similar. The training speed (convergence of loss) was almost the same. However, the training itself on pure skeletons was noticeably faster than on heatmaps and grayscale images. Skeletons take up much less memory space than images, resulting in fewer computations. Despite the fact that their accuracy is slightly lower, using skeletons instead of heatmaps is reasonable as it speeds up the training process without sacrificing much accuracy.

TABLE III: Top-1 accuracy results for different models on validation with added noise with $\alpha = 0.015$

Learning Type	Model	Input Type	Model label	alpha	gamma_1	gamma_2	sup_tresh	Top-1 Acc
Contrastive	PoseConv3D	Skeleton	C_Sk_1	0.01	1	0.3	0	74.7
			C_Sk_2	0.025	5	1	0	80.2
		3D Heatmap	C_H_3	0.01	1	0.3	20	57.4
	MotionBERT	Skeleton	C_Sk_MB_4	0.01	1	0.3	0	71.5
Supervised	PoseConv3D	Skeleton	S_Sk_5	-				53.5
		3D Heatmaps	S_H_6					5.1
		Grayscale	S_G_7					8.6

B. *Contrastive learning*

For contrastive learning experiments, we used the PoseConv3D and MotionBERT models as backbones. The experiments were conducted with different values of γ_1 and γ_2 (coefficients for instance contrastive loss and group contrastive loss, respectively), as well as different values of the sup_thresh - parameter, which determines the number of initial epochs with supervised training only (without contrastive loss). We also investigated the effect of the α parameter, which represents the percentage of noise in the data, specifically the percentage of the original video size by which we shifted the initial keypoint.

To assess the reliability, we validated the obtained models on noisy validation data. We selected several values of the α parameter (percentage of noise in the data):

```
alphas = [0, 0.001, 0.005,
          0.01, 0.015, 0.02,
          0.03, 0.05, 0.1]
```

and added noise to the original validation data in a similar manner to the training data: shifting each point according to a zero-mean Gaussian distribution. The results of the best-performing models are presented in Figure 2 and Table III.

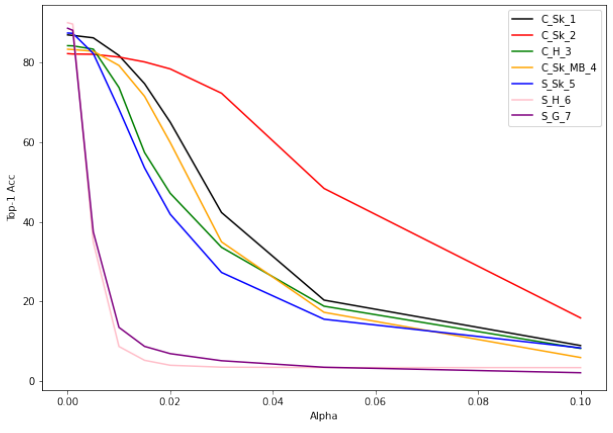

Fig. 2: Validation results for the models described in Table III for different alpha

We observe that supervised learning significantly lags behind contrastive learning in terms of robustness. This is particularly noticeable for the two lower models: 3D Heatmaps and Grayscale (`S_H_6` and `S_G_7`). This indicates that the initial PoseConv3D model is especially unstable. However, PoseConv3D trained on skeletons (`S_Sk_5`) achieves much better results, although it still falls behind contrastive learning. Such a difference in robustness to noise between training on images and pure skeletons may be attributed to the fact that heatmaps lose some spatial information due to skeleton interpolation and transformation. On the other hand, training on pure skeletons involves learning from precise initial coordinates, preserving fine-grained spatial details that can be crucial for noise robustness.

It is also interesting to note that PoseConv3D and MotionBERT exhibit almost parallel behavior for the same α parameter (`C_Sk_1` and `C_Sk_MB_4`), indicating that the models have nearly identical noise robustness.

Furthermore, for larger values of α (`C_Sk_2`), PoseConv3D may experience a decrease in accuracy on validation data without noise, but it demonstrates greater robustness to strong noise.

V. Conclusion

In this study, we explored the contrastive learning method for introducing noise into the data to obtain more robust models. We improved state-of-the-art PoseConv3D and MotionBERT models in the task of action recognition and addressed some of their limitations, particularly enhancing noise robustness. The obtained models showed significant improvements on noisy data compared to the original models, indicating that contrastive learning is applicable for generating robust models.

Furthermore, we investigated different types of input data for the PoseConv3D model and found that the model performs much faster on raw skeletons without sacrificing much accuracy.

Further work could explore the robustness of the obtained models on smaller data volumes. For example, it would be interesting to examine how well the models cope with noise when trained on only 5%, 10%, or 20% of the data. Moreover, experimentation with the nature of the added noise can be conducted. In our study, we used Gaussian noise. However, training for other noise distributions could lead to even more robust models.

REFERENCES

[1] Haodong Duan, Yue Zhao, Kai Chen, Dahua Lin, and Bo Dai. Revisiting skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2969–2978, 2022.

[2] Ankit Singh, Omprakash Chakraborty, Ashutosh Varshney, Rameswar Panda, Rogerio Feris, Kate Saenko, and Abir Das. Semi-supervised action recognition with temporal contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10389–10399, 2021.

[3] Wentao Zhu, Xiaoxuan Ma, Zhaoyang Liu, Libin Liu, Wayne Wu, and Yizhou Wang. Learning human motion representations: A unified perspective. *arXiv preprint arXiv:2210.06551*, 2022.

[4] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

[5] J. Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. pages 4724–4733, 07 2017.

[6] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Learning spatio-temporal features with 3d residual networks for action recognition. pages 3154–3160, 10 2017.

[7] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang 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*, pages 1010–1019, 2016.