MSTCN-VAE: An unsupervised learning method for micro gesture recognition based on skeleton modality

WenXuan Yuan
2799782134@qq.com

Taiyuan university of technology

August 21,2023

Catalogue

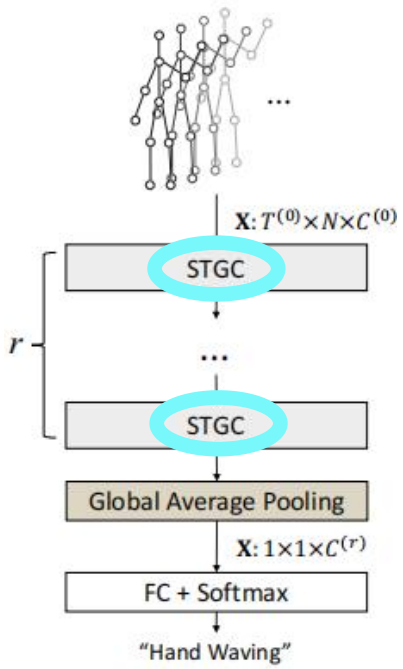
Catalogue

Introduction & Related Work

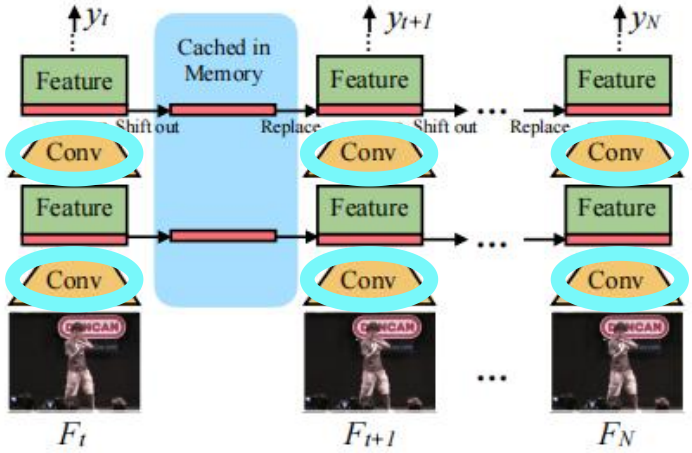
- Recently, Skeleton-based approaches offer a compact and informative representation that captures the spatial and temporal dynamics of human actions, enabling the efficient processing of gestures and facilitating the extraction of relevant features for recognition tasks.
- While skeleton-based action recognition has achieved remarkable success, there is a growing interest in exploring micro-gesture recognition, which focuses on recognizing subtle and fine-grained hand movements.

Introduction & Related Work

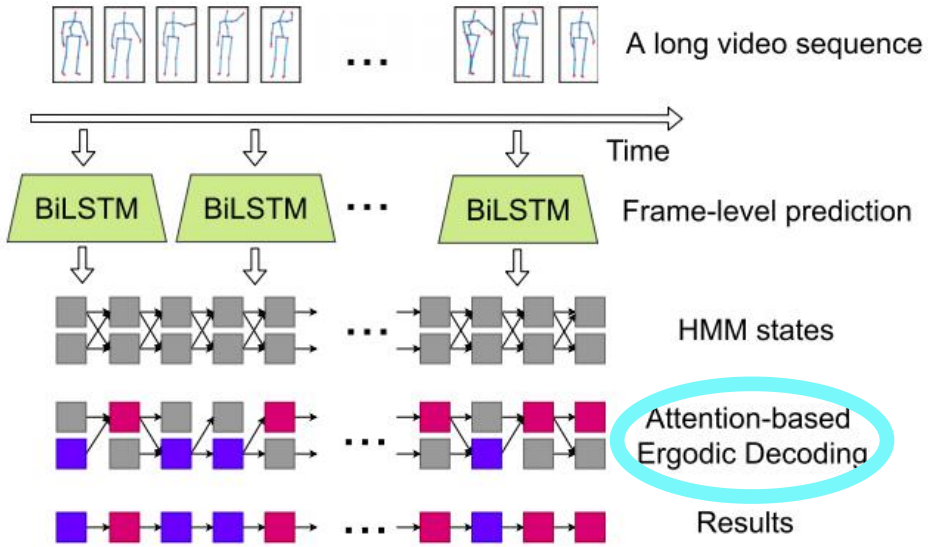
- To address this research frontier, many methods have been proposed.



MS-G3D model



TSM model



HMM model

Introduction & Related Work

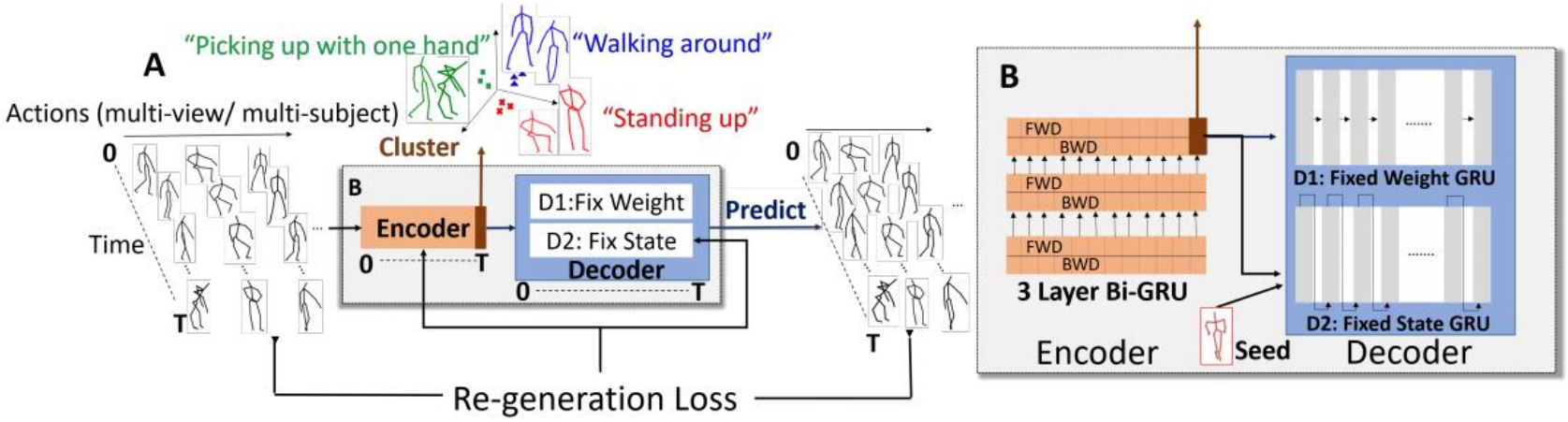
- Although there are already many excellent supervised Skeleton-based methods, these approaches rely on labels that we have m-ade.
- If we take a supervised approach, we must classify this set of act-ions into the types of actions we already know, and there may be some kinds we can't discern.

Introduction & Related Work

- In order to overcome the dependence of supervised methods on labels and the difficulty of labeling micro-gesture dataset, researchers have proposed some innovative unsupervised methods
  - Predict & Cluster(P & C)
  - U-S-VAE

Introduction & Related Work

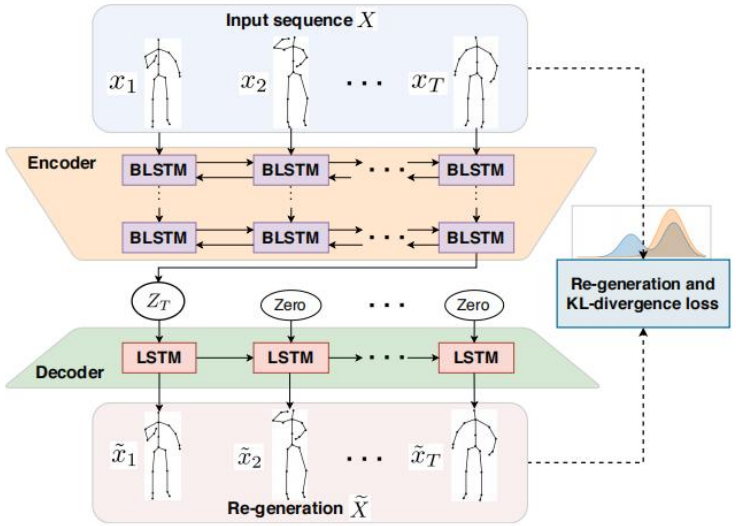
- P & C based on an encoder-decoder system. P&C provides a way to automatically recognize actions from skeletal data through multi-layer bidirectional GRU and shows promising results on NW-UCLA, UW-A3D and NTU RGB-D 60 datasets.



P & C model

Introduction & Related Work

- In U-S-VAE, the experimental results show the effectiveness of using multi-layer bidirectional LSTM to extract human skeleton information in the iMiGUE dataset.



U-S-VAE model

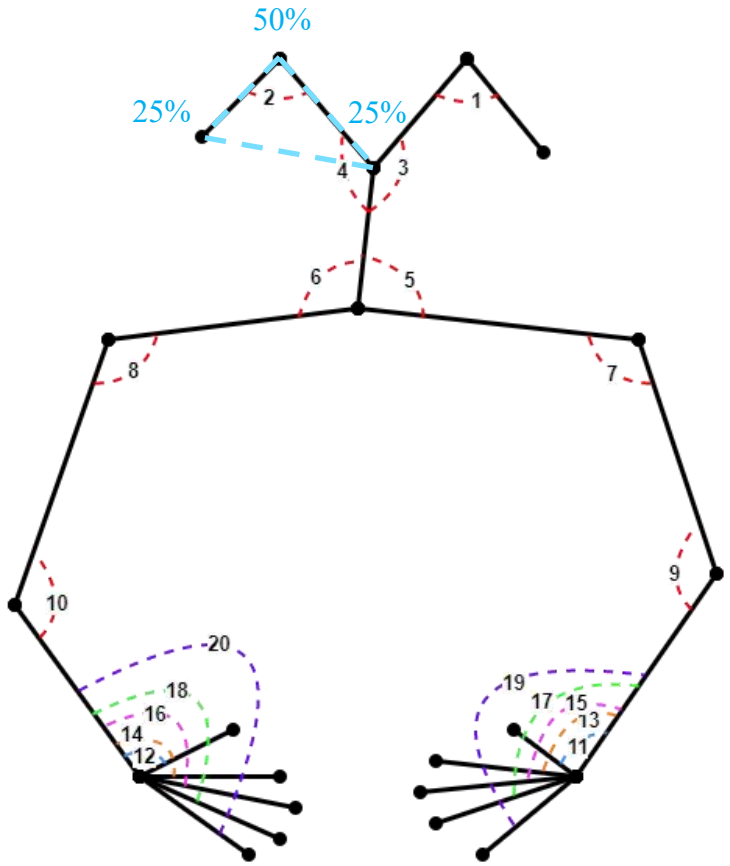
Introduction & Related Work

- U-S-VAE and P&C both use the encoder-decoder network structure. However, TCN and MSTCN have been shown to be better than RNN, LSTM and GRU in terms of time integration in a variety of scenarios.
- Further, our model continues to use encoder-decoder structure, However, we respectively embed TCN and MSTCN in the encoder and put deconvolutional neural network into the decoder.

Catalogue

Preliminary Data Processing

- We calculated the angle between the connection lines of the joint nodes in each frame in terms of the right figure.
- The confidence degree is the weighted average of the three points of a corner, 25%, 50%, 25% respectively.



Preliminary Data Processing

- Although the Angle information extracted in this way ignores the length between the joint nodes with certain importance, it directly reduces the data dimension.
- Moreover, for subtle action sequences, such as micro-gestures, small changes in Angle are more indicative of the meaning of the action itself than the length.

Preliminary Data Processing

- This constitutes a two-dimensional Angle information Extracted data(AE data).
- While the data in the dataset that has not been changed in any w-ay, i.e. Original Skeleton data, will be referred to as OS data.

	Sample	Channel	Frame length	Joint / Angle	People in each frame
AE data	same	2	same	20	same
OS data	same	3	same	22	same

Table: Comparison of AE data and OS data

Catalogue

Model Architecture

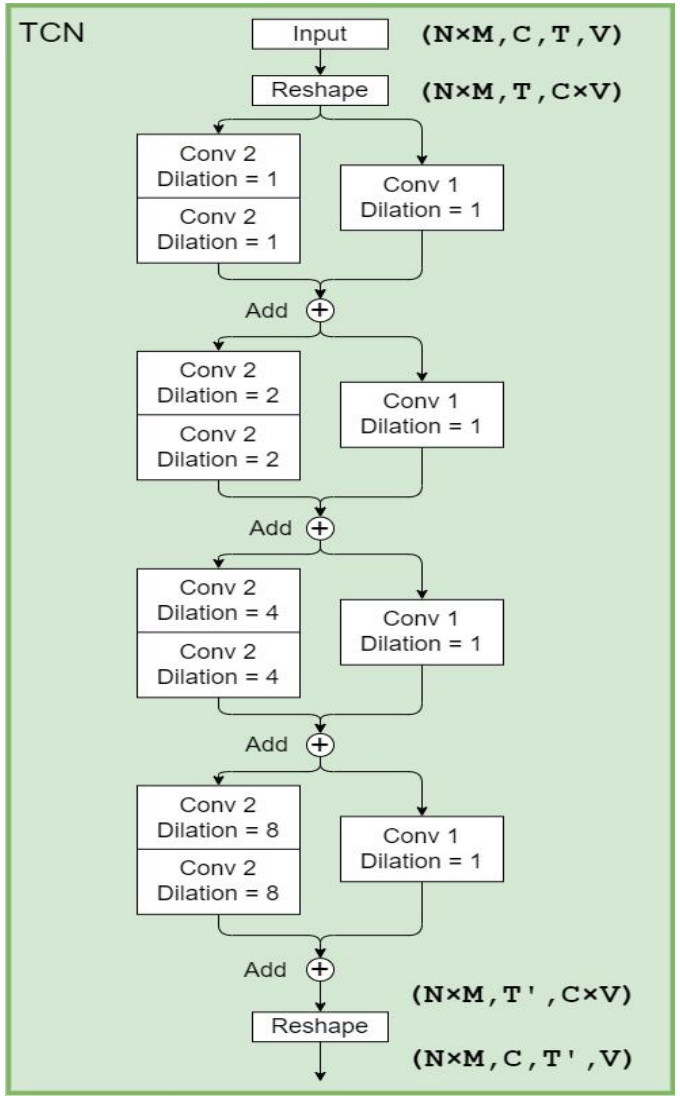
- Notation Statement
  - N: Number of samples
  - C: Number of channels
  - T: Frame length after downsampling
  - V: Number of joints
  - M: The number of people in each frame

Model Architecture

- Temporal Convolutional Network (TCN) block

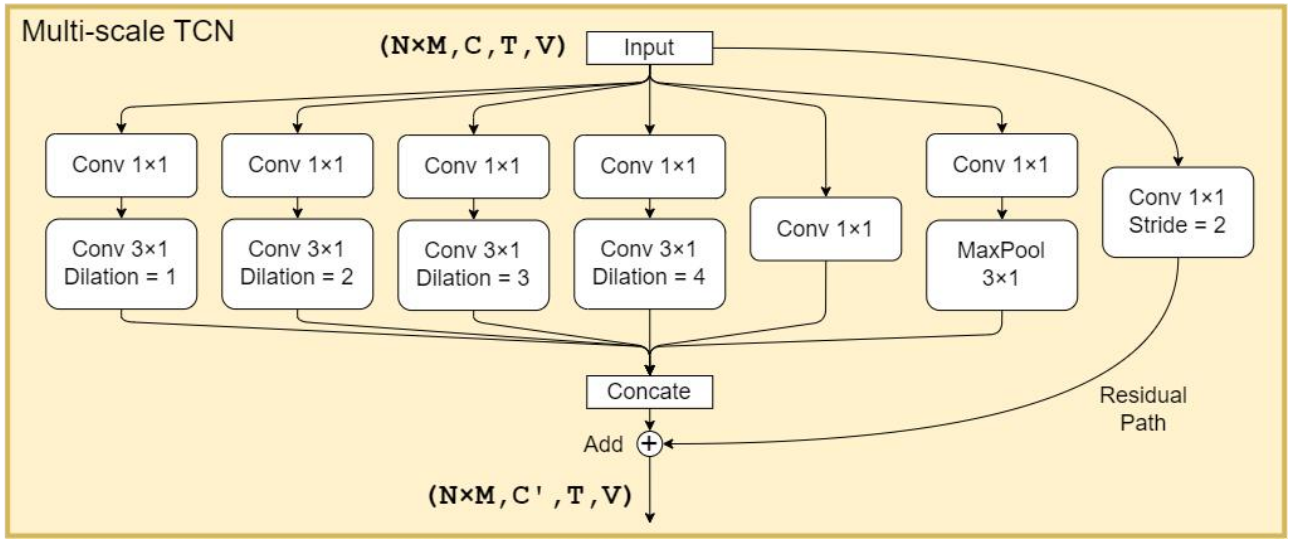
  Firstly, the input four-dimensional data is r-eshaped, undergoes multiple convolutions i-n different dilations, and then reshaped agai-n into four-dimensional data in preparation for the hidden feature extraction block.

  - **Details are shown on the right**



Model Architecture

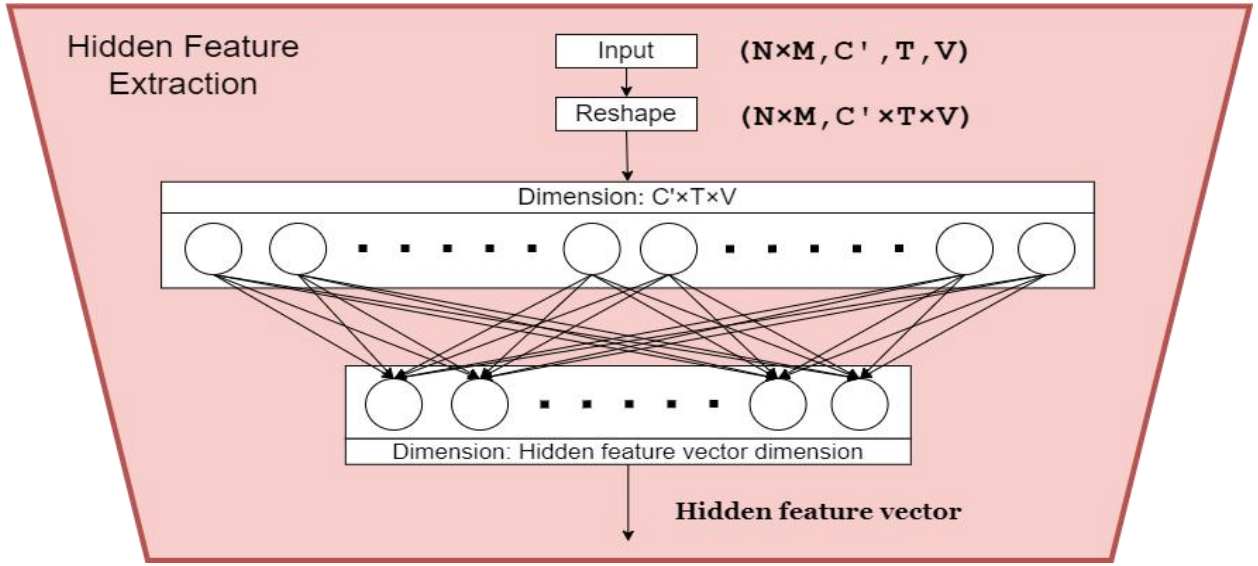
- Multi-scale TCN (MSTCN) block
  - This block is the same as the configuration of MS-TCN block in MS-G3D model.
  - **Details are shown on the below**



- Hidden Feature Extraction(HFE) block

  The data, which is temporally integrated by TCN or MSTCN, is fed into a fully connected neural network to obtain the hidden feature vector.

  - Details are shown on the below

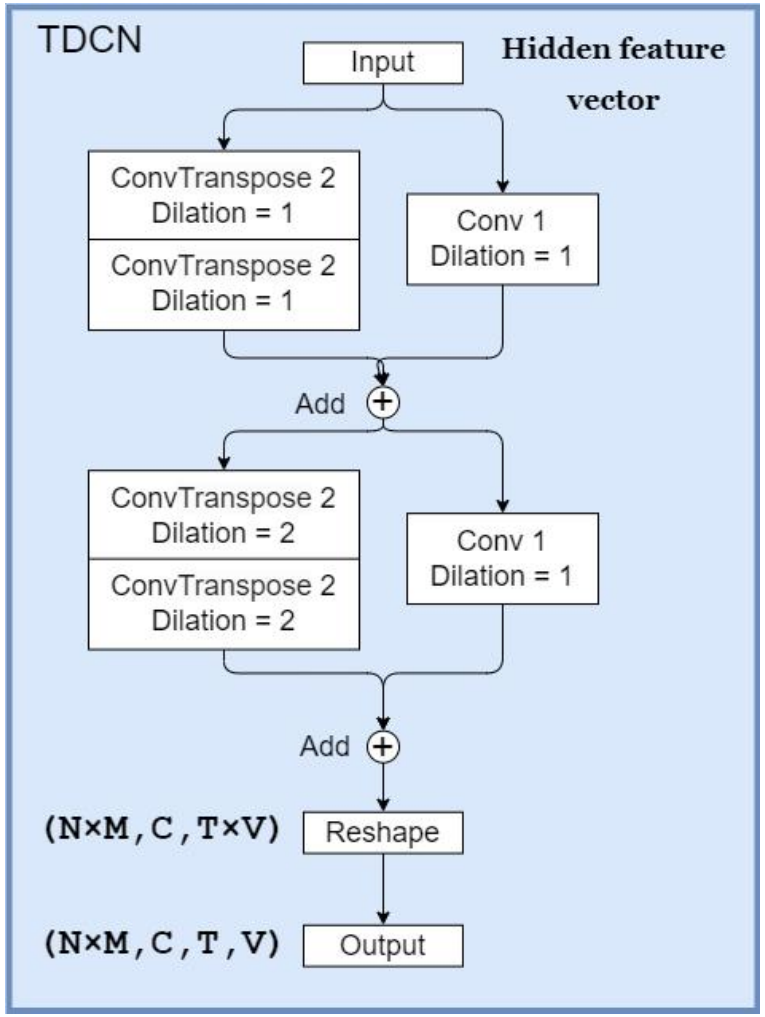


Model Architecture

- Temporal Deconvolutional Network (TDCN) block

  The hidden feature vector is deconvolved several times on the time dimension and r-eshaped as the output.

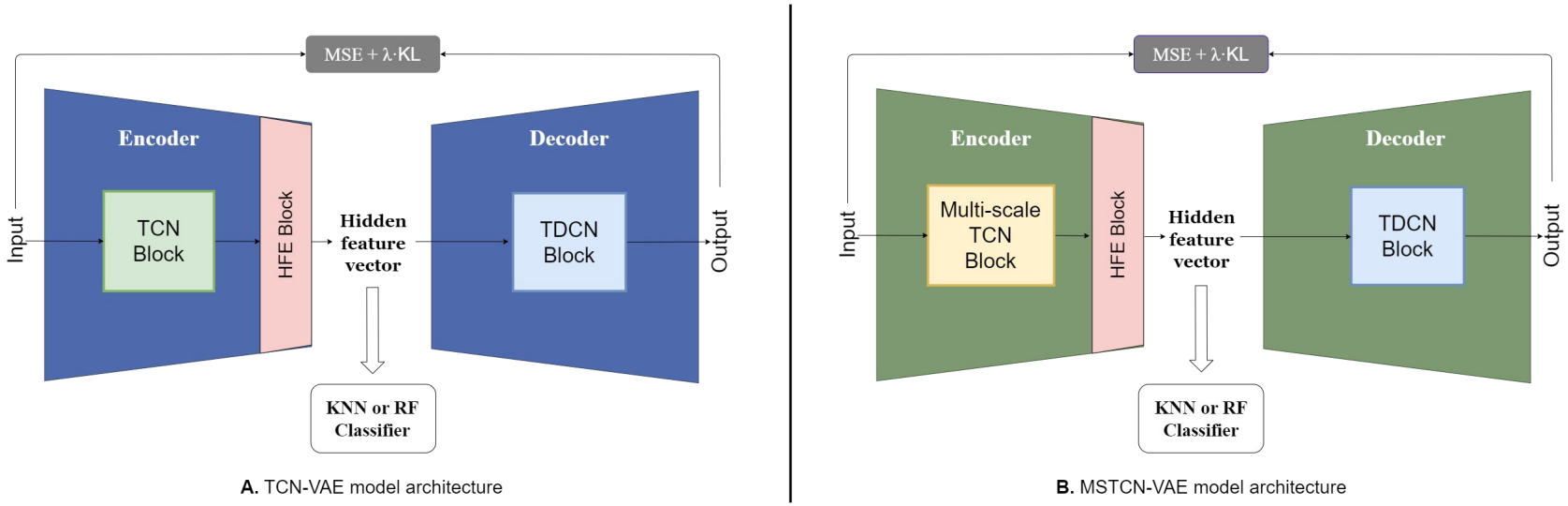
  - **Details are shown on the right**



VAE Architecture

The encoder is composed of TCN or MSTCN block and HFE block, and the decoder is embedded with TDCN block. The loss function consists of MSE + $\lambda \cdot$ KL divergence, and we found that $\lambda = 8$ is a dandy parameter value.

- Details are shown on the below



A. TCN-VAE model architecture

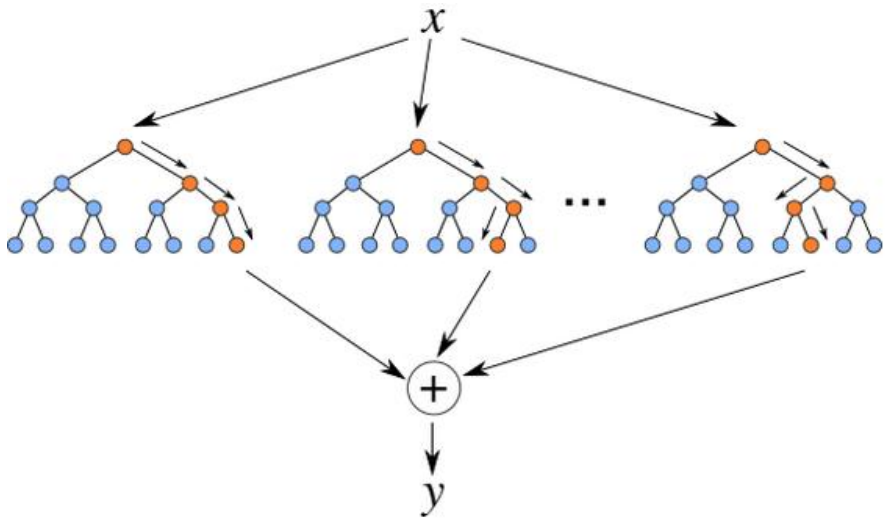
B. MSTCN-VAE model architecture

- ## Unsupervised Classification

  When calculating the accuracy, all the sequence data in the training set are forward propagated in the current network to obtain the hidden feature vectors of all the training data, and this is used to form the KNN classification space. After the same forward propagation for each sample in the test set, the KD-tree algorithm is used to quickly search for the neighboring samples in the just-formed classification space.

Supervised Classification

The supervised process is similar to the unsupervised process, except we replace the KNN classifier with the Random Forest classifier (RF classifier). It is important to clarify that since the RF classifier uses the label information to form the classification space, the model with this classifier belongs to the supervised network.



Random Forest

Classification Details

In the model evaluation session, for our different MSTCN-VAE variants:

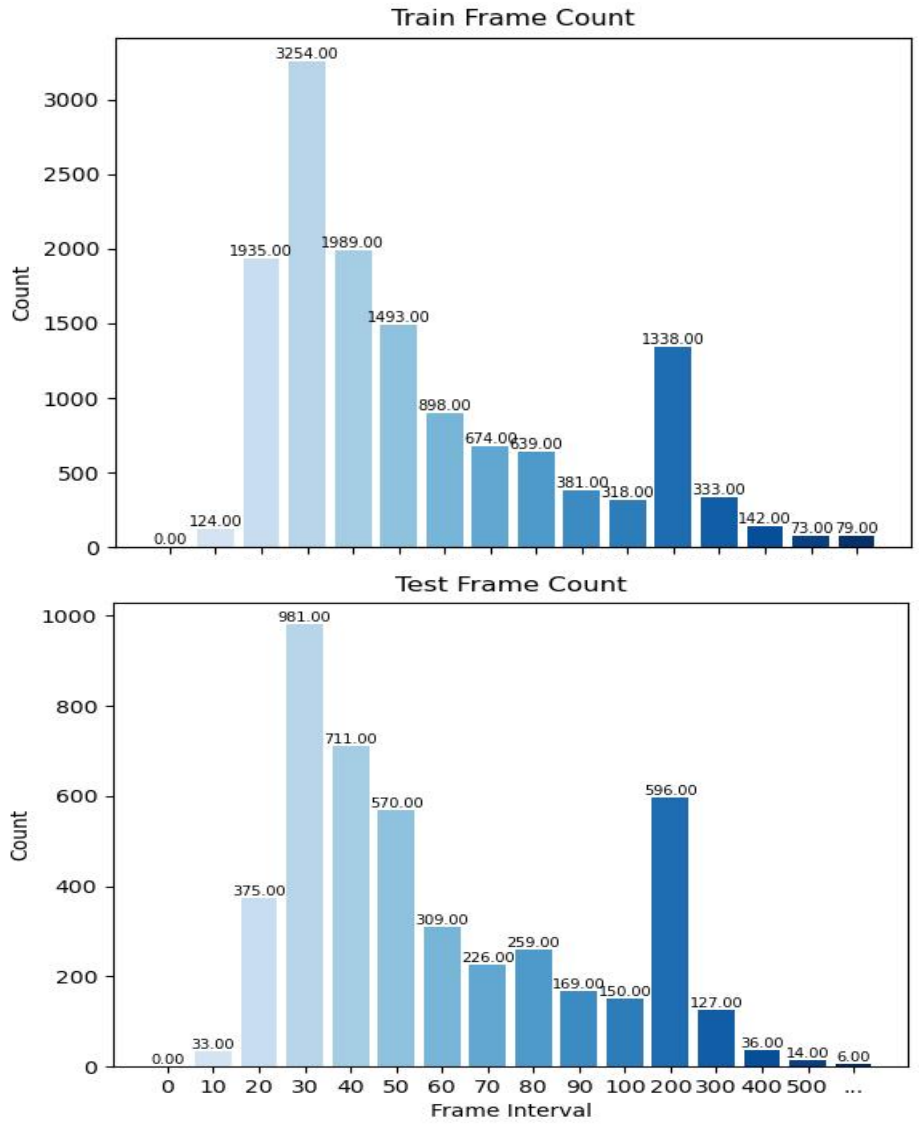
- **Unsupervised:** Top1 accuracy and Top5 accuracy were calculated using k=1 and k=1,2,3,4,5 under the KNN classifier.
- **Supervised:** Top1 accuracy and Top5 accuracy were calculated using random_state =1 and random_state =1, 2, 3, 4, 5 under the RF classifier.

In the Top5 calculation, for the classification results under five different parameters of the classifier, the prediction is considered correct as long as it contains the correct category.

Catalogue

Implementation Details

- To train the network, each action sample was downsampled by up to 100 frames. T-he joint point data in each skeleton map were also normalized.
- For the optimization hyperparameters
  - optimizer: SGD
  - batch size: 32
  - initial LR: 0.0001
  - max_epoch: 200
  - LR decay percent: 10%
  - LR decay step: (100, 150).



Implementation Details

In all the models (TCN-VAE and MSTCN-VAE), the blocks are designed as follows:

- TCN block, convolves T-dimension into 75, 50, 25, and 1.
- TDCN block, deconvolutes T-dimension of sizes 50 and 100.
- MSTCN block, similar to the setting for MS-TCN block in MS-G3D.

Implementation Details

- For the models without HFE block, the hidden feature vectors used for classification are all 66-dimensional, and for the models using HFE block, the hidden feature vectors are 128-dimensional.
- In order to avoid gradient explosion during the training process, gradient truncation will be performed when the maximum norm is greater than 10.

iMiGUE dataset	SMG dataset
AE data + OS data	OS data

Experiment

iMiGUE dataset			
	Methods	Top1	Top5
Unsupervised	P&C	31.67	64.93
	U-S-VAE	32.43	64.30
	TCN VAE (with out HFE) (OS data) (**Our**)	24.44	39.54
	TCN VAE (OS data) (**Our**)	28.50	44.91
	MSTCN VAE (OS data) (**Our**)	30.84	45.22
	MSTCN VAE (AE data+OS data) (**Our**)	35.38	50.07

Experiment

iMiGUE dataset			
	Methods	Top1	Top5
Supervised	S-VAE	27.38	60.44
	ST-GCN	46.97	84.09
	Shift-GCN	51.51	88.18
	MS_G3D	54.91	89.98
	TCN_VAE(with RF classifier) (OS data) (**Our**)	39.11	48.55
	MSTCN_VAE(with RF classifier) (OS data) (**Our**)	41.23	51.64
	MSTCN_VAE(with RF classifier) (AE data) (**Our**)	35.73	45.20
	MSTCN_VAE(with RF classifier) (AE data + OS data) (**Our**)	47.69	56.36

Experiment

SMG dataset			
	Methods	Top1	Top5
Supervised	ST-GCN	41.48	86.07
	shift-GCN	55.31	87.34
	MS_G3D	64.75	91.48
	MSTCN_VAE (with RF classifier) (OS data) (**Our**)	42.59	49.54
Unsupervised	MSTCN_VAE (OS data) (**Our**)	30.06	45.28

Catalogue

Conclusion

- We propose a skeleton-based micro-gesture recognition method. Our model connects a MSTCN network with a HFE block as an encoder to aggregate out hidden feature vectors and uses a temporal deconvolutional network in the decoder to generate action sequences from the hidden feature vectors.
- Through experiments on the iMiGUE dataset, we develop and demonstrate the improvement of the MSTCN-VAE model over previous unsupervised methods, in addition to validation on the SMG dataset to further illustrate the effectiveness of our model.

Conclusion

- Because of our team was inexperienced. This model does not take the gr-aph connectivity property into a-ccount in the VAE structure.
- Moreover, for the application of our model, I think it should not be lim-ited to skeleton data only. Introducing multi-modal data will provide c-omplementary information to each other.

Thanks For Listening

WenXuan Yuan
2799782134@qq.com

Taiyuan university of technology

August 21,2023