CV PROJECT

- Team ID: 15

- Team Members:
    - Sai Jashwanth (20171178), Sai Soorya Rao (20171052), Raviteja (20171067)
    - TA Mentor : Pranay Gupta

- Project ID, TITLE : 19, **View Adaptive Neural Networks for High Performance Skeleton-based Human Action Recognition**

ABSTRACT

- One of the key challenges in action recognition lies in the large variations of action representations when they are captured from different viewpoints. This paper introduces a novel view adaptation scheme, which automatically determines the virtual observation viewpoints over the course of an action in a learning based data driven manner.

- They designed two view adaptive neural networks, i.e., VA-RNN and VA-CNN. For each network, a novel view adaptation module learns and determines the most suitable observation viewpoints, and transforms the skeletons to those viewpoints for the end-to-end recognition with a main classification network.

PROBLEM REASONING

What are the Problems faced before:

There are two major reasons for large view variations:

- ❖ First, in a practical scenario, the viewpoints of the cameras are flexible and different viewpoints result in large differences in skeleton representations even for the same scene.
- ❖ Second, the actor could conduct an action in different orientations. Moreover, he/she may dynamically change his/her orientations as time goes on.

Problem with view-invariant transformation pre-processing, used in previous works:

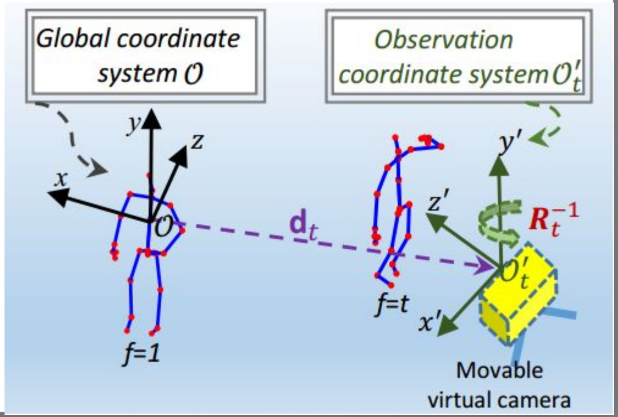
- ❖ For **Frame-level pre-processing**, where each frame is transformed to the body center with the upper body orientation aligned, usually results in the partial loss of relative motion information.
- ❖ For **Sequence-level pre-processing**, in which case the motion is invariant to the initial body position and orientation ,and the motion information is preserved. However, since the human body is non-rigid, the definition of the body plane by the joints of “hip”, “shoulder”, “neck” is not always suitable for the purpose of orientation alignment.

PROBLEM FORMULATION

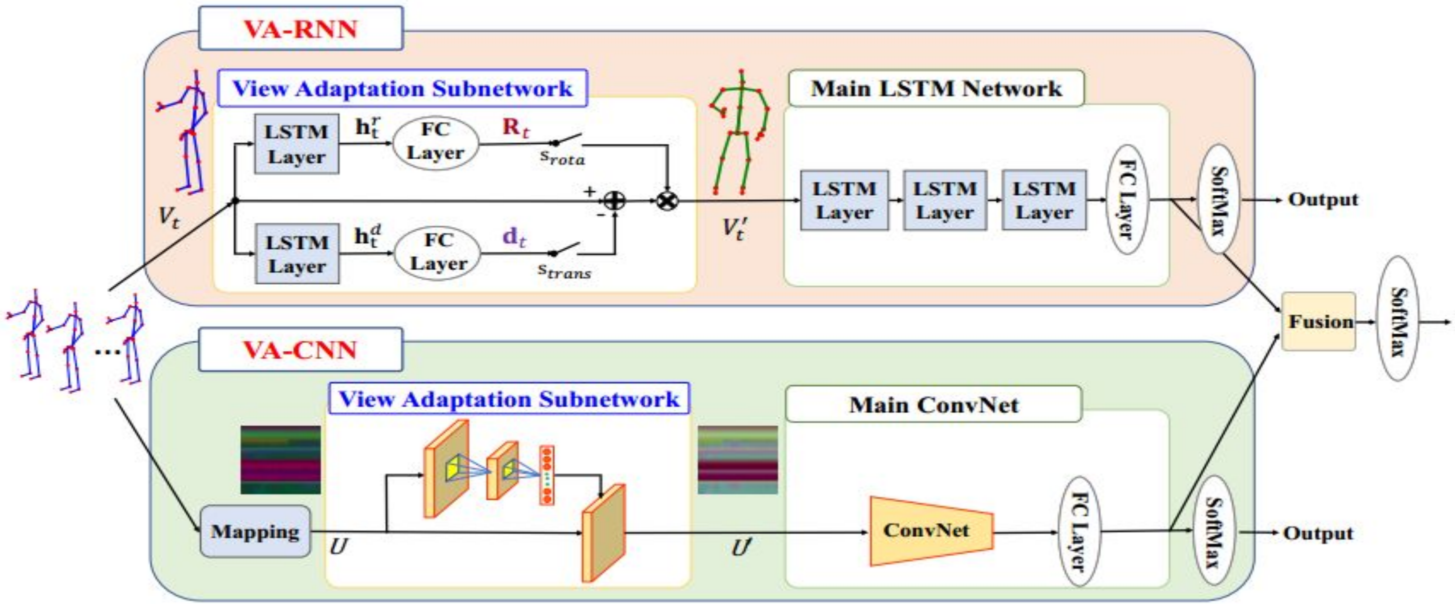
Given a skeleton sequence S under the global coordinate system O, the jth skeleton joint on the t'th frame is denoted as $\mathbf{v}_{t,j} = [x_{t,j}, y_{t,j}, z_{t,j}]^{\mathrm{T}}$, where $t \in (1, \cdots, T)$, T denotes the total number of frames a sequence, $j \in (1, \cdots, J)$, J denotes the total number of skeleton joints in a frame. The set of joints in the t'th frame is denoted as $V_t = \{\mathbf{v}_{t,1}, \cdots, \mathbf{v}_{t,J}\}$.

$$\mathbf{v}'_{t,j} = [x'_{t,j}, y'_{t,j}, z'_{t,j}]^{\mathrm{T}} = \mathbf{R}_t(\mathbf{v}_{t,j} - \mathbf{d}_t). \quad \mathbf{R}_t = \mathbf{R}^x_{t,\alpha}\mathbf{R}^y_{t,\beta}\mathbf{R}^z_{t,\gamma},$$

The skeleton representation, $V'_t = \{\mathbf{v}'_{t,1}, \cdots, \mathbf{v}'_{t,J}\}$ under new observation coordinate.



WORKFLOW


VA-RNN
View Adaptation Subnetwork
LSTM Layer
FC Layer
LSTM Layer
FC Layer
V_t
V'_t
Main LSTM Network
LSTM Layer
LSTM Layer
LSTM Layer
FC Layer
SoftMax
Output
Fusion
SoftMax
VA-CNN
View Adaptation Subnetwork
Mapping
U
U'
Main ConvNet
ConvNet
FC Layer
SoftMax
Output

View Adaptive Convolution Neural Network (VA-CNN)

- First we Map **Skeletons to Image,** with columns representing different frames while rows representing different joints. The 3D coordinate values for X, Y ,and Z are treated as the three channels of an image.

$$\mathbf{u_{t,j}} = floor(255 \times \frac{\mathbf{v_{t,j}} - \mathbf{c_{min}}}{c_{max} - c_{min}}),$$

- Next **View Adaptation Subnetwork:**
  - **Main ConvNet:** Transformed skeleton map as input, we can use an existing ConvNet, for classification.

$$\begin{aligned}\mathbf{u'_{t,j}} &= 255 \times \frac{\mathbf{v'}_{t,j} - \mathbf{c_{min}}}{c_{max} - c_{min}} \\ &= \mathbf{R}_{t,j}\mathbf{u}_{t,j} + 255 \times \frac{\mathbf{R}_{t,j}(\mathbf{c}_{min} - \mathbf{d}_{t,j}) - \mathbf{c}_{min}}{c_{max} - c_{min}}.\end{aligned}$$

View Adaptive Recurrent Neural Network (VA-RNN)

View Adaptation Subnetwork:

➢ At a time slot corresponding to the t'th frame, with a skeleton Vt as input, two branches of LSTM subnetworks are utilized to learn the rotation matrix Rt, and the translation vector dt.

- The branch of rotation subnetwork for learning rotation parameters consists of an LSTM layer, and a fully connected (FC) layer.

$$[\alpha_t, \beta_t, \gamma_t]^{\mathrm{T}} = \mathbf{W}_r \mathbf{h}_t^r + \mathbf{b}_r,$$

- The branch of translation subnetwork for learning translation parameters consists of an LSTM layer, and a FC layer. $\mathbf{d}_t = \mathbf{W}_d \mathbf{h}_t^d + \mathbf{b}_d,$

Main LSTM Network: The LSTM network is capable of modeling long-term temporal dynamics and automatically learning feature representations. The number of neurons of the FC layer is equal to the number of action classes. Softmax Classifier is used.

Training: Let us denote the loss back propagated to the output of the view adaptation subnetwork as $\epsilon_{v'_{t,j}} \in \mathbb{R}^{1\times 3},$

Loss back-propagated to branch for translation, rotation parameters.

$$\epsilon_{\mathbf{d}_t} = -J\epsilon_{v'_{t,j}} \mathbf{R}_t, \quad \epsilon_{\alpha_t} = \epsilon_{v'_{t,j}} \frac{\partial \mathbf{R}_t}{\partial \alpha_t} \sum_{j=1}^{j=J} (\mathbf{v}_{t,j} - \mathbf{d}_t).$$

GOALS

- Creating DataSet in addition to available dataset(NTU RGB-D) by performing view enriching by rotation of the skeleton around the axes during training procedure.
- Implementing the VA-CNN for VA-Subnetwork.
- Implementing the VA subnetwork first and then implementing the main classifier network.
- Obtaining and comparing the results with that of paper.
- Implementing VA-RNN for course project if time permits.

ESTIMATED TIMELINE

WORK	DATE
Implementing VA CNN subnet	20-02-2020
Implementing Main ConvNet	02-03-2020
MID EVALUATION	05-03-2020
Creating view enriched data set	10-03-2020
End-to-End Training on Dataset	20-03-2020
Final Submission	28-03-2020