SRM INSTITUTE OF SCIENCE AND TECHNOLOGY
FACULTY OF ENGINEERING AND TECHNOLOGY
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

A translation medium for Text to avatar-based Indian sign language translation by synthesizing motion capture data using Deep Learning

Presented by :
Sowmya J Iyer (RA1711003010034)
T.S. Meghna (RA1711003010098)

Guided by: Ms. P. Saranya, Assistant Professor (Sr.G)

SRM
INSTITUTE OF SCIENCE & TECHNOLOGY
(Deemed to be University u/s 3 of UGC Act, 1956)

Abstract

- This project aims at leveraging the challenge of using 3D poses for Sign Language translation or animation by transforming 2D pose datasets into 3D ones.Automation of this process is valuable because the manual method causes time overheads.
- The goal is, using a 3D dataset of American Sign Language, to train a deep neural network that will predict the depth coordinates of the skeleton key points from 2D coordinates
- Animating a person in 3D graphics requires a huge set up with motion trackers to track the person's movements and also takes time to animate each limb manually. We aim to provide a time-saving method to do the same.
- We use Open Pose for 2d pose estimation and use DeepSORT+FaceReID for motion tracking in RGB Videos
- Various methods such as LSTM, deep CNNs and GANs are explored for 2D to 3D rendering
- Rendering to 3D: The coordinates of these 17 landmark points detected in the previous step will now be the positions of the joints of limbs of the 3D character required to be animated.

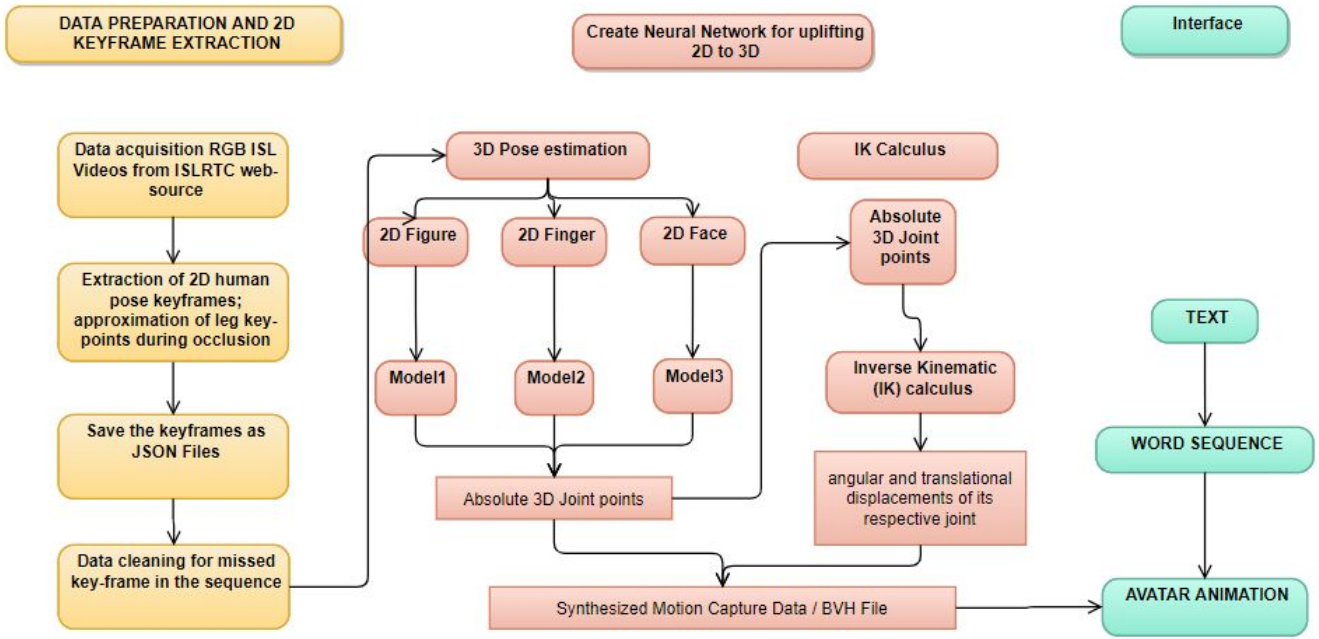
SRM
INSTITUTE OF SCIENCE & TECHNOLOGY
(Deemed to be University u/s 3 of UGC Act, 1956)

Objective

- To present a novel pipeline for the generation of 3D skeleton data from sign videos for subsequent use as data augmentation tool in sign translation scenarios.
- To transfer the 3D data to animate an avatar that interprets texts and audios.

SRM INSTITUTE OF SCIENCE & TECHNOLOGY (Deemed to be University u/s 3 of UGC Act, 1956)

Architecture Diagram



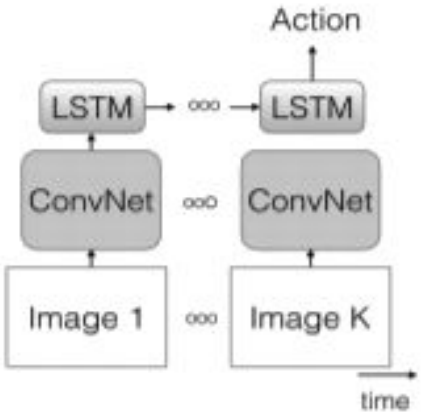
SRM
INSTITUTE OF SCIENCE & TECHNOLOGY
(Deemed to be University u/s 3 of UGC Act, 1956)

Module Description & Implementation

Module 1: Data Extraction and 2D keyframe extraction.

- ISLRTC videos are scraped using the pytube library and then these are used for the development and the testing part of the model.
- Further, DeepSort+ FaceReID are used to track the movements and build a smooth animation
- OpenPose is used to extract the 2D keyframe of the human pose, all the 17 points which include- hand, body and face.
- Akima interpolation is used to get rid of the missing data issue as it gave a natural and smooth curve while fixing the data.

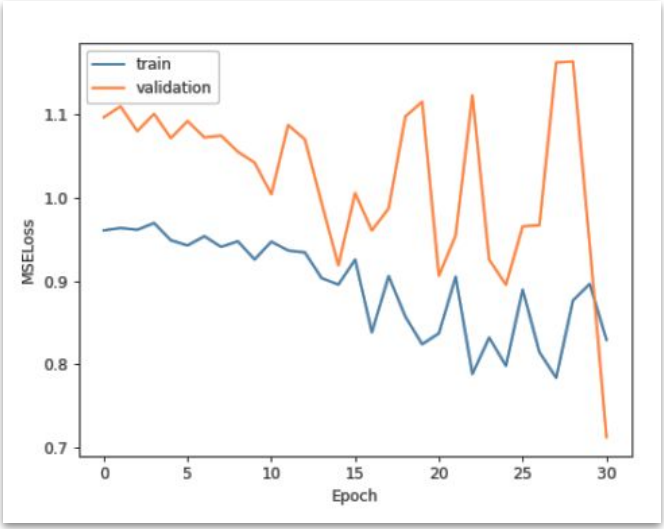
Module 2: UPLIFTING 2D TO 3D KEYFRAME USING HOURGLASS NETWORK

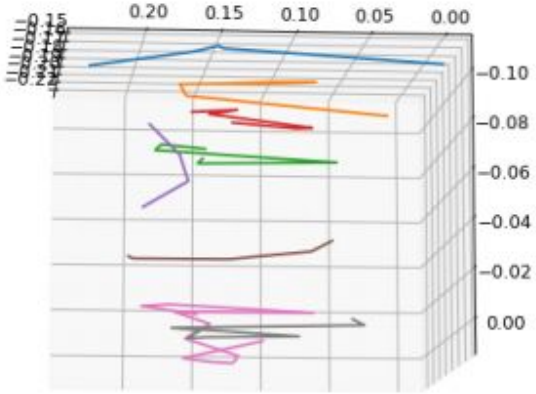


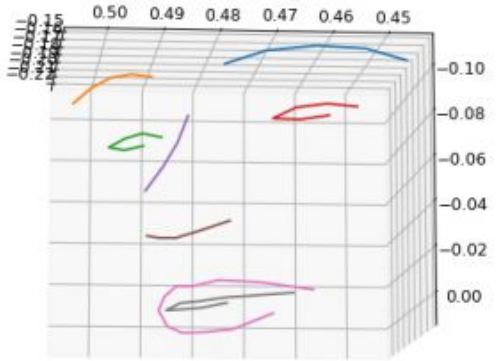
MODEL1 : ConvNET + LSTM

- LSTM - handles the encoding of state and the temporal ordering. Also, dependencies with long-ranges are captured.
- A LSTM layer was equipped with batch normalization (as proposed by Cooijmans et al.) after the last average pooling layer of Inception-V1, with 512 hidden units. Above the classifier, a fully connected layer is attached. This model was attempted to be trained using cross-entropy losses on the outputs at all time steps.
- The original video was 42 frames-per-second stream. Subsampling of input video frames were done by keeping one out of every 7 frames.
- a LSTM with N-layers followed by a fully connected layer with ReLU activation is built.
- Tested on three variations of this architecture.

Performance of LSTM





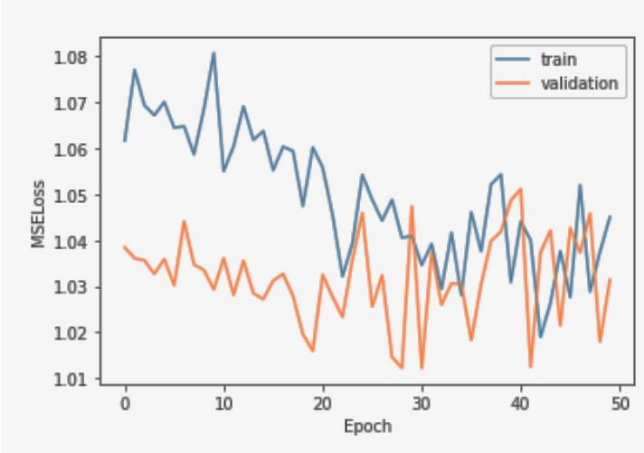


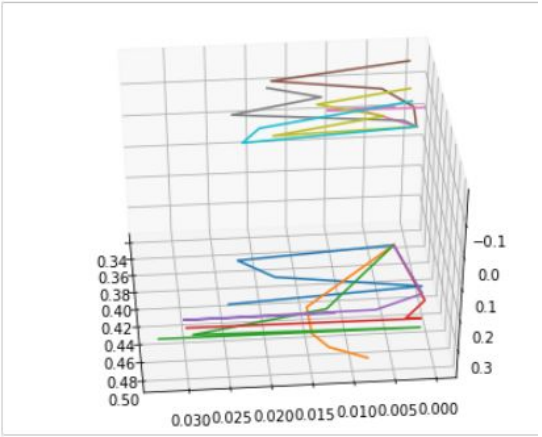
VARIATION 1: LSTM with 2 layers and 512 units used as hidden units

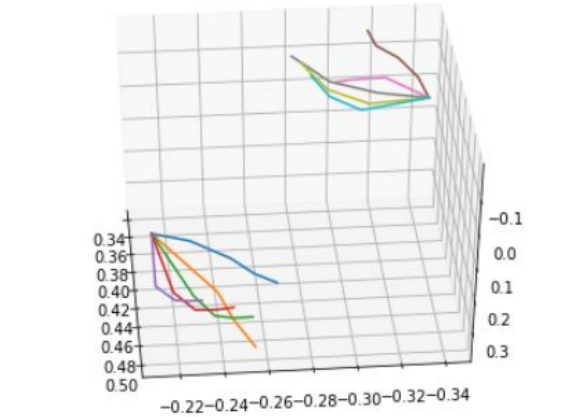
Test loss: 1.5886

Performance : VERY POOR

Performance of LSTM





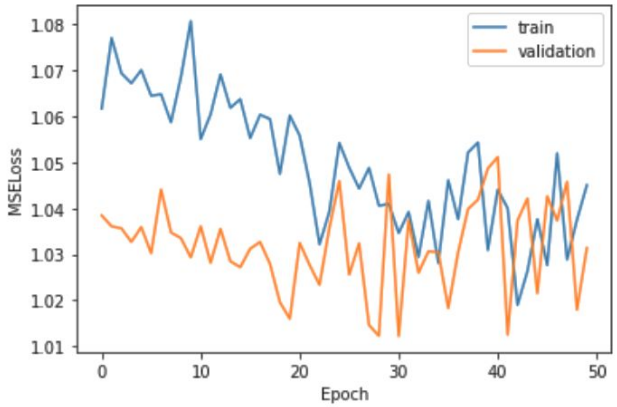


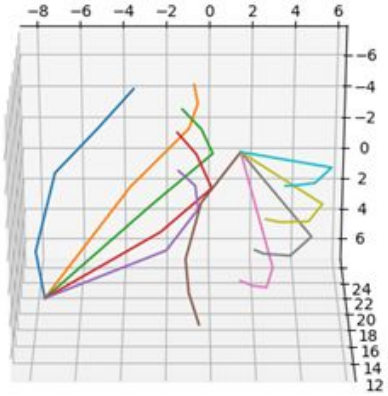
VARIATION 2: bidirectional LSTM and 512 units used as hidden units (8,644,634 trainable parameters)

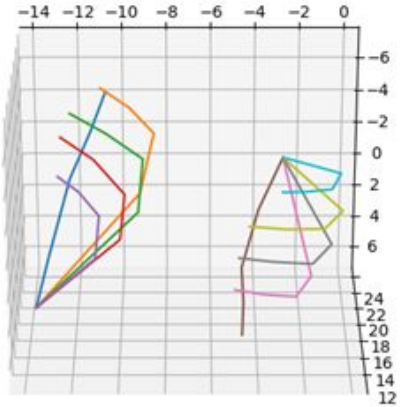
Test loss: 1.5886

Performance : VERY POOR

Performance of LSTM





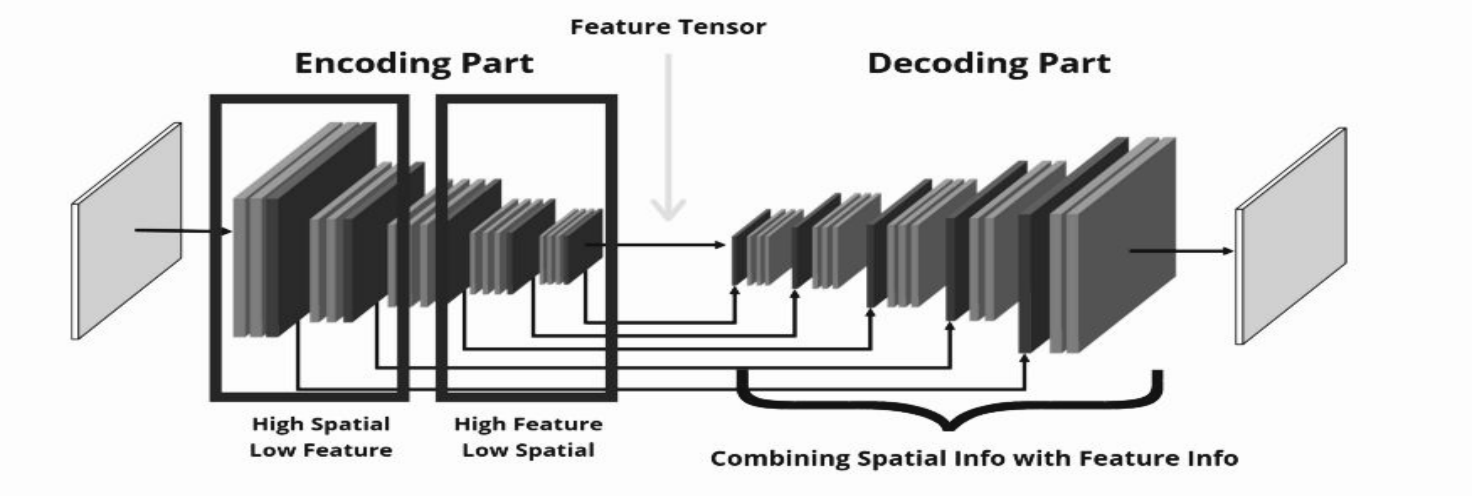


VARIATION 3: bidirectional LSTM and 1024 units used as hidden units (34,066,458 trainable parameters)

Test loss: 0.7562

Performance : UNSATISFACTORY

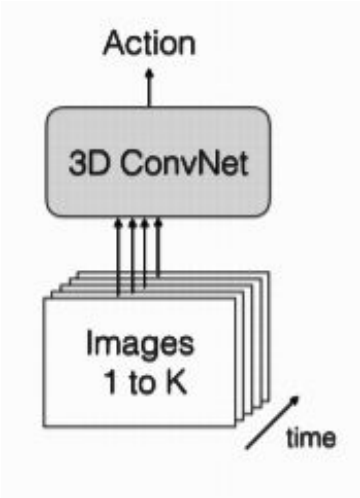
3D CNN based stacked Hourglass network



MODEL 2 : Hourglass Network

- Highly similar to models which are equipped with processing spatial information for depth prediction such as fully convolutional networks and ResNETs.
- Has a symmetric distribution of capacity between bottom-up processing (high resolutions to low resolutions) and top-down processing (low resolutions to high resolutions).
- Instead of unpooling/ deconvolution layer, nearest neighbour upsampling is used.
- Connections of top-down processing is skipped

Parts of the Network - Network Architecture



3D ConvNETS: Feature Extraction

- **As conventional as standard convolutional networks. Additionally contain spatio-temporal filters that directly create hierarchical data representation in the spatial temporal region.**

Inflating 2D ConvNets into 3D.

- **Instead of building 3D convNETS from scratch, by simply converting state-of-the-art 2D pose models into 3D ConvNets we obtain a high performance spatio-temporal models. By starting with the 2D hourglass architecture, all the filters and pooling kernels are inflated to give a z-value as temporal.**

Max Pooling Layers: Eliminate parts that are not important for feature extraction

Residual Layers: Push Layers deeper into the network using skip layers.

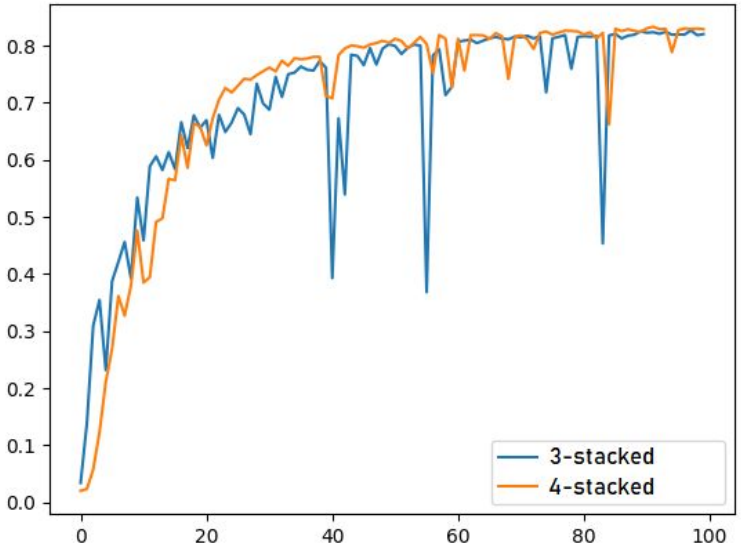
Bottleneck Layers: to lessen the complexity of the network.

- ResNets use residuals heavily throughout the network. They are used to combine the spatial info with the feature info.
- Here, we use a bottleneck block which are a new type of residual. This saves a lot of memory by implementing the use of 2 1x1 convolution layers instead of a 3x3 layer. Making the computation speed lesser and computations easier.

Stacked HourGlass:

- multiple hourglasses are stacked end2end
- Reiteration and reconfirming of initial estimates and features across the neighboring frames
- intermediate hourglass predictions are fused back into the feature space
- This is done by mapping them to a larger number of channels along with an additional 1x1 convolution.
- It is important to note that weights are not shared across hourglass modules, and a loss is applied to the predictions of all hourglasses using the same ground truth.

Current Results over 15 videos



Validation score

Optimisation : rmsprop

Learning rate: 2.5e-4

GPU: 12 GB NVIDIA TitanX GPU

Training time : 3 days

A single forward pass: 75ms

Validation score : 81.625

Currently, drops after 40 epochs

Module 3:TESTING AND DEPLOYMENT

- We import the .bhv (Bio vision Hierarchy Data) file and the attractive, user friendly avatar into the blender.
- imported data is present in the form of individual animation clips, which makes every empty independent of each other and provides us with the feature to modify any element without disturbing the other
- the animation is baked, in this process, all the independent animations are baked together into a single piece of motion for a smooth transition.
- some manual adjustments are made according to the output we get and we bake the animation again to find the final output.

Contributions

- 3D data from videos has been extensively looked at from the action recognition application of Computer vision domain.
- The idea is to implement 3D CNN inflation for the above to obtain human pose estimation from videos.
- The inflation of the 2D convolutions to 3D convolutions is used to perform 3D human pose estimation in spatial temporal domain.
- Features from neighbouring frames are learnt through this inflation and further used to refine its predictions.
- We found an increase from 64.8 to 63.6 in MPJPE (Mean Per Joint Position Error) through this inflation method

SRM
INSTITUTE OF SCIENCE & TECHNOLOGY
(Deemed to be University u/s 3 of UGC Act, 1956)

Results and Discussion

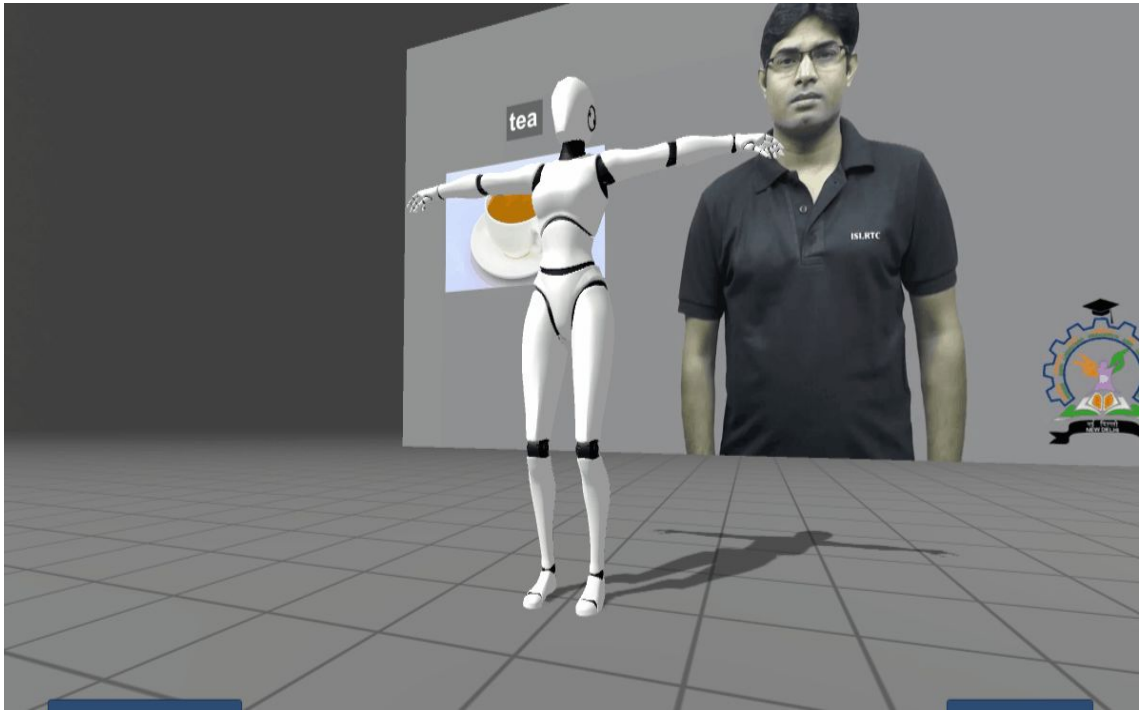
- Video stabilization provided better and helped in acquiring stable, non-shaky smooth videos.
- JSON file is converted to .csv file for ease of data access.
- Akima interpolation predicted nature of missing joints more dynamically than linear interpolation.
- **LSTM model is discarded due to poor performance when compared to hourglass network**
- The unwanted joints are removed from the avatar, in order to provide a smooth animation result.
- The .bhv files feeded into the avatar only contained key points of the upper body. Hence. Stagnant key-points are provided to the thighs and the legs to get the desired results.
- Manual adjustments are done in the rig of the avatar to fit it properly to the .bhv file data.

SRM
INSTITUTE OF SCIENCE & TECHNOLOGY
(Deemed to be University u/s 3 of UGC Act, 1956)

Screenshots

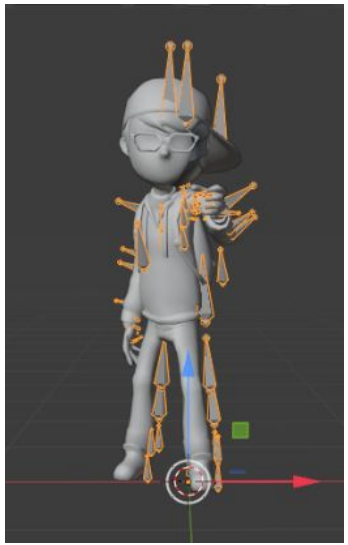
Module 1:

Room for Improvement



- Approximation of occluded joint by normalising the key frames
- Suppressing 3D model utilisation for faster training
- Way to make the model use lesser resources.

Module 3:



SRM
INSTITUTE OF SCIENCE & TECHNOLOGY
(Deemed to be University u/s 3 of UGC Act, 1956)

References

1] Disabled Persons in India - A statistical Profile 2016
http://mospi.nic.in/sites/default/files/publication_reports/Disabled_persons_in_India_2016.pdf

[2] Havasi, L., & Szabo, H. M. (2005). A Motion Capture System for Sign Language Synthesis: Overview and Related Issues. EUROCON 2005 - The International Conference on “Computer as a Tool.” doi:10.1109/eurcon.2005.1629959

[3] Bugra Tekin, Pablo Márquez-Neila, Mathieu Salzmann, and Pascal Fua. 2017. Learning to fuse 2d and 3d image cues for monocular body pose estimation. In Proceedings of the IEEE International Conference on Computer Vision. 3941–3950.

[4] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2017. OpenPose: Realtime multi-person 2D pose estimation using part affinity fields. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 7291–7299.

[5] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. 2019. MediaPipe: A framework for building perception pipelines. arXiv preprint arXiv:1906.08172 (2019).

[6] F. Zhang, X. Zhu, H. Dai, M. Ye, and C. Zhu, "DistributionAware Coordinate Representation for Human Pose Estimation," CVPR, vol. abs/1910.06278, 2019.

[7] S Ertürk, TJ Dennis, Image sequence stabilization based on DFT filtering. IEE Proc. Vision Image Signal Process. 147(2), 95–102 (2000).

8] C Song, H Zhao, W Jing, Y Bi, in Proc. Intl. Conf. Pattern Recognition. Robust video stabilization based on bounded path planning, (2012).

[9] M Grundmann, V Kwatra, I Essa, in Proc. IEEE Conf. Computer Vision and Pattern Recognition. Auto-directed video stabilization with robust l1 optimal camera paths, (2011).

[10] Filling the Gaps: Predicting Missing Joints of Human Poses Using Denoising Autoencoders Computer Vision – ECCV 2018 Workshops, 2019, Volume 11130

[11] Two-dimensional video-based analysis of human gait using pose estimation Jan Stenum, Cristina Rossi, Ryan T. Roemmich bioRxiv 2020.07.24.218776;doi: https://doi.org/10.1101/2020.07.24.218776

[12] M. Trajkovic and M. Hedley ´ , Fast corner detection, Image and Vision Computing, 16 (1998), pp. 75–87. https://doi.org/10.1016/S0262-8856(97)00056-5.

[13] L. Moisan, P. Moulon, and P. Monasse, Automatic Homographic Registration of a Pair of Images, with A Contrario Elimination of Outliers, Image Processing On Line, 2 (2012), pp. 56–73. http://dx.doi.org/10.5201/ipol.2012.mmm-oh.

[14] B.D. Lucas and T. Kanade, An iterative image registration technique with an application to stereo vision, in Proceedings of the 7th International Joint Conference on Artificial intelligence, vol. 81, 1981, pp. 674–679.

[15] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, “Realtime multi-person 2d pose estimationusing part affinity fields,” in CVPR, 2017

[16] Two-dimensional video-based analysis of human gait using pose estimationJan Stenum, Cristina Rossi, Ryan T. Roemmich bioRxiv 2020.07.24.218776;doi: https://doi.org/10.1101/2020.07.24.218776