# Declaration

I hereby declare that the work presented in this Thesis titled *"Assessment of Human Actions in Videos"* submitted to the Indian Institute of Technology Jodhpur in partial fulfilment of the requirements for the award of the degree of Doctor of Philosophy, is a bonafide record of the research work carried out under the supervision of *Dr. Gaurav Harit*. The contents of this thesis in full or in parts, have not been submitted to, and will not be submitted by me to, any other Institute or University in India or abroad for the award of any degree or diploma.

*Hiteshi Jain*
*P14CS001*

# Certificate

This is to certify that the thesis titled *"Assessment of Human Actions in Videos"*, submitted by *Hiteshi Jain (P14CS001)* to the Indian Institute of Technology Jodhpur for the award of the degree of *Doctor of Philosophy*, is a bonafide record of the research work done by her under my supervision. To the best of my knowledge, the contents of this report, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.

*Dr. Gaurav Harit*
*Ph.D.Thesis Supervisor*

# Abstract

Humans have the desire to achieve better performance and outcome of their efforts. In performances such as sports, yoga or professions such as surgery, humans strive to attain perfection and efficiency. They improve their performances by comparing themselves with others or seeking feedback from experts. In this regard the question arises here; can we make a human action assessment system which can support the users with a flexibility of using the system at preferred time and place, thereby avoiding the dependency on a trainer.

The key challenge in action quality assessment is the lack of an appropriate metric as it is quite a subjective task with a significant influence of human (expert) bias. The labels like scores or skill levels provided by human judges lack interpretability. Thus, it will be interesting to develop an automated action assessment system that can bring more interpretability and objectivity to this domain. Action assessment system can be made more objective by considering a set of reference high-precision/expert-level action videos. In this thesis, we propose techniques to assess human actions, where the problem of action assessment is transformed into the problem of comparing a given action video with a reference video. We consider two use-cases : Yoga and Sports, and assess actions like Sun Salutation, Diving and Gymnastic Vaults.

We contribute a new Sun Salutation Assessment dataset that includes expert and non-expert performances and their respective ground truth judgments. Sun Salutation is a long term complex action with many possible errors and is a good test-bed for human action assessment techniques. Physical exercises like Sun Salutation and Aerobics are repetitive in nature. Such exercises require the performer to achieve consistent dynamic poses and smooth transitions. The quality check for such exercises involves the analysis of attributes such as pace, consistency, smoothness, etc. of the performers. We propose a framework to identify jerky and inconsistent movements in a performance that involves action segmentation using Hidden Markov Model, followed by an inter-pose timing analysis.

Based on the performance skill, humans can be divided into experts, mid-level, and amateurs. Amateur performers are prone to miss or wrongly perform a part of some action. We propose a template matching-based approach where the pose sequence of the test performer is compared with an expert sequence and the action segments that were missed or anomalously performed are reported. Template-based matching is a fairly simple approach to identify differences and has difficulty in scaling when the number of templates tends to grow. In practice, actions can be performed correctly in multiples ways. Thus there is a need for comparison of the test sequence to all expert renderings possible for an action sequence. Towards this, we propose an unsupervised sequence-to-sequence autoencoder-based model that learns to reconstruct all expert videos. The skill level of a test performance is judged based on how well the learned model can reconstruct the test sequence. The closer the test performance is to an expert, the more accurately it gets reconstructed.

Actions like diving or gymnastic vaults are constrained with limited availability of expert performances, which makes it difficult to train an autoencoder-based assessment model. To address this limitation, we propose a Deep Metric Learning based framework, where a Long Short Term Memory (LSTM)-based Siamese network learns to predict if two videos in a pair are similar or dissimilar based on the difference in their scores. The learned model is utilized for action scoring where the performances are compared with the reference expert performance to determine the score.

...

# Acknowledgments

This work has been done during the great time I have spent at Indian Institute of Technology Jodhpur. I would like to express my gratitude to my advisor, Dr. *Gaurav Harit* for introducing me to this exciting field of human action analysis and for his endless support and encouragement and for providing me the appropriate resources and fruitful discussions on the topic.

Thank you to all faculties who have groomed me in different ways both in academics and with the different roles and responsibilities.

My sincere gratitude to *Tata Consultancy Services* for sponsoring my Ph.D. under the TCS Research Scholarship Program and for bearing my expenses for the conferences.

This work would not be the same without the support from all my friends at IIT Jodhpur: *Brajesh Shukla, Rakesh Kumar, Venu Acche, Parmod Paul, Pradumn Pandey, Vivek Tiwari, Divya Sharma, Megha Singh, Arpita Jaitawat, Divya Gupta, Manju Kumari, Deepak Dhillon and Bhuvnesh Rathore*. Thanks for the great atmosphere and fun.

Finally, I would like to thank *Onkar Krishna, Charu Goyal, my parents, my sister and family* for their infinite patience and care and for boosting my confidence throughout this span.

*Hiteshi Jain*
*Ph.D. Student*

# List of Figures


| *Figure* | *Title* | *page* |
|---|---|---|
| 1.1 | Organization of Thesis | 4 |
| 2.1 | a) Outline contour of a walking subject and their corresponding dot configuration b) motion paths of seven side joints of the walking person (reprinted from [Johansson [1973]] © Springer, 1973) | 11 |
| 2.2 | 3-D model description of humans using cylindrical parts (reprinted from [Marr and Nishihara [1978]] © The Royal Society, 1978) | 11 |
| 2.3 | a) Kinect camera with its camera configuration b) Pose estimation using a depth camera (reprinted from [Shotton *et al.* [2013]] © IEEE, 2013). Body part label at each pixel is predicted using Body part classification (BPC) and used to localize joints | 12 |
| 2.4 | Mixture-of-parts model [Yang and Ramanan [2011]; Felzenszwalb and Huttenlocher [2005]]; in the top-right model a single template is warped to different orientations. In the bottom-right, the body parts are Approximated using small warps by translating patches connected with a spring (reprinted from [Yang and Ramanan [2011]] © IEEE, 2011) | 13 |
| 2.5 | Mesh feature (reprinted from [Yamato *et al.* [1992a]] © IEEE, 1992), sample shape masks for foreground stroke from tennis | 15 |
| 2.6 | Sample MEI and MHI images (reprinted from [Bobick and Davis [2001]] © IEEE, 2001) | 15 |
| 2.7 | Stacked silhouettes from action frames (reprinted from [Blank *et al.* [2005]], © IEEE, 2007) | 15 |
| 2.8 | a)Shape context feature for human representation (reprinted from [Belongie *et al.* [2002]], © IEEE, 2002), b)matrix representation of shape context. Reference silhouette with 2 query silhouettes | 16 |
| 2.9 | Space time interest point features for a walking sequence (reprinted from [Laptev [2005a]], © Springer, 2005) | 16 |
| 2.10 | 2D vs 3D convnet models (reprinted from [Tran *et al.* [2015]], © IEEE, 2015) | 17 |
| 2.11 | 3D pseudo resnet : an alternate to 3D convolution with separate spatial and temporal filters (reprinted from [Qiu *et al.* [2017]], © IEEE, 2017) | 18 |
| 4.1 | Sun Salutation Sequence | 32 |
| 4.2 | Timeline illustrating jerks, rest times and consistency. From top to bottom sequence 1) 2 cycles of fast Sun Salutation by Yoga expert with no jerk, no rests, and consistent action. 2) 1 cycle of slow Sun Salutation by Yoga expert without jerk and uniform rest times 3) 1 cycle of slow Sun Salutation with jerk and non-uniform rest times 4) 2 cycles of inconsistent Sun Salutation with jerks | 33 |
| 4.3 | Mesh features for a posture of Sun Salutation | 34 |
| 4.4 | Sample frames of Sun Salutation with variable no. of STIP features | 35 |
| 4.5 | Concatenated Hidden Markov Model for long-term action sequence | 37 |
| 4.6 | Block Diagram for Fast/Slow Sun Salutation decoding | 37 |
| 4.7 | Key poses of Action 4 and 7; Red arrows indicate motion direction of body parts | 38 |
| 4.8 | Activity graph of a Sun Salutation sequence | 40 |
| 4.9 | Illustration of false alarms and missed anomaly detections | 42 |
| 5.1 | Pose-vector representation | 46 |
| 5.2 | Approximate String Matching illustration | 47 |

5.3 DTW alignment: (a) with missed actions (b) anomalous actions 49
5.4 Misalignments due to Dynamic Time Warping 49
5.5 Anomalous Poses in Amateur Videos 50
5.6 False Alarm for anomalous segments due to incorrect pose estimation 51

6.1 Sequence-to-Sequence Autoencoder model to learn temporal evolution of expert videos 54
6.2 Scoring of Test Sequences 56
6.3 Rank correlation of individual template videos 57
6.4 Reconstruction Results on some sample videos. Top to Bottom: First bar shows the ground truth sun salutation sequence followed by five test sequences and their respective reconstructed videos. Different actions are missed in different videos. The reconstructed sequence matches the ground truth sequence in every case.(Different colors denote different poses. Black color denotes padding used to generate fixed length video) 58

7.1 Our community detection based segmentation technique is unsupervised and requires no training and can segment without knowledge of count of action categories 62
7.2 Analysis of community detection results for video dataset 65
7.3 a) Shape context feature for Weizmann dataset human silhouette b) MHAD101-s human motion skeletal data 67
7.4 Qualitative segmentation results where *GT* : Ground Truth; Different colors correspond to distinct actions 68
7.5 Comparative results for all videos 68
7.6 Anomalous pose detection result 69
7.7 Incorrect sequences of the dataset; $A$ denotes the anomalous actions 70
7.8 Segmentation results : GT refers to the ground truth; different colors represent different poses; Gray color indicates transition frames between the key poses 71

8.1 Siamese Architecture for Deep Metric Learning 77
8.2 Score Estimation Relative to Expert 78
8.3 Clip level similarity Estimation for evaluating individual clip contribution 80
8.4 Architecture variations 84
8.5 Shows the t-SNE plots of LSTM embedding and Siamese embedding of training and test data. Column 1 represents the distribution of embeddings of all the data points with different colours of different dive types. Columns 2,3,4 represent the score-wise distribution of 3 distinct dive types where the radii of the circles drawn are proportionate to the score of the corresponding embedding. Row1 shows the plots of the LSTM embedding $O_q$. Subsequent rows show Siamese embeddings at $D_2$ with different reference biases used at $O_p$. Row2 shows the plots when using a Uniform Bias at $O_p$. Rows 3, 4 and 5 show plots with dive type specific biases applied at $O_p$. Row3 uses the worst performer bias, Row4 uses the 50 percentile score performer bias, Row5 uses the 75 percentile score performer bias and Row6 uses the expert bias. 90

# List of Tables


| *Table* | *Title* | *page* |
|---|---|---|
| 4.1 | Performance accuracy of posture specific HMM | 43 |
| 4.2 | Comparison of segmentation accuracy of concatenated HMM with traditional Viterbi and the proposed Viterbi Algorithm | 43 |
| 4.3 | Performance accuracy of assessment framework in detecting jerks and inconsistent segments | 43 |
| 5.1 | Performance accuracy for missed action notification (under tolerance of 25 frames) | 51 |
| 5.2 | Performance accuracy for anomalous action notification (under tolerance of 25 frames) | 51 |
| 5.3 | Number of Edit Operations of different test sequences with 5 templates of same expert | 51 |
| 6.1 | Comparison of Rank Correlation and Mean Square Error of various techniques for Sun Salutation performance scoring | 59 |
| 7.1 | Segmentation Accuracy : Community based approach outperforms even without knowledge of no. of actions | 69 |
| 7.2 | Comparison of Rank Correlation and Mean Square Error of various techniques for Sun Salutation action scoring with Community Detection based encoding | 70 |
| 7.3 | Segmentation results : Sun Salutation video segmentation in the presence of incorrect performances | 71 |
| 8.1 | Data Splits. P: Positive pairs, N: Negative Pairs | 81 |
| 8.2 | UNLV Diving Results - Comparison of Rank Correlation; C:C3D, L:LSTM, F: Final, I: Incremental, S: SVR. Dataset splits are as described in Section 8.5(A). | 83 |
| 8.3 | UNLV Diving Results - Comparison of Mean Square Error; C:C3D, L:LSTM, F: Final, I: Incremental, S: SVR. | 83 |
| 8.4 | Results of our model in comparison to the baseline works for various dive types. Dive-type specific results are averaged over the splits described in Section 8.5(A) | 83 |
| 8.5 | UNLV Gymnastic Vaults - Comparison of Rank Correlation and Mean Square Error; C:C3D, L:LSTM, F: Final, I: Incremental, S: SVR, RGR: Reference Guided Regression | 84 |
| 8.6 | Multitask AQA Dataset - Comparison of Rank Correlation and Mean Square Error; C:C3D, L:LSTM, F: Final, I: Incremental, S: SVR, RGR: Reference Guided Regression | 84 |
| 8.7 | Impact of Metric Learning on final scoring | 86 |
| 8.8 | Architecture variations | 86 |
| 8.9 | Clip Level contribution to the final score | 88 |
| 8.10 | Performance of feedback module on original videos and when the videos are misaligned; RGR_s$X$ denotes the reference guided regression performed when we shift the candidate videos with $X$ frames to create misalignments | 89 |
| 8.11 | Performance comparison our model with D2 embedding generated by including worst bias (WB), bias from performer of 50 and 75 percentile and expert bias (EB) | 89 |
| 8.12 | Runtime of different methods | 91 |

# List of Abbreviations

| Abbreviation | Full form |
|---|---|
| *ACA* | Aligned Cluster Analysis |
| *AQA* | Action Quality Assessment |
| *ASM* | Approximate String Matching |
| *C3D* | Convolution 3D |
| *CD* | Community Detection |
| *CNN* | 2-Dimensional |
| *CRF* | Conditional Random Field |
| *DCT* | Discrete Cosine Transform |
| *DFT* | Discrete Fourier Transform |
| *DTW* | Dynamic Time Warping |
| *HACA* | Hierarchical Aligned Cluster Analysis |
| *HAR* | Human Activity Recognition |
| *GT* | Ground Truth |
| *GRU* | Gated Recurrent Unit |
| *HMM* | Hidden Markov Model |
| *LDA* | Latent Dirchlet Allocation |
| *LSTM* | Long Short Term Memory |
| *pLSA* | Probabilistic Latent Semantic Analysis |
| *P3D* | Pseudo-3D Resnet |
| *RNN* | Recurrent Neural Network |
| *S3D* | Segment-based P3D-fused network |
| *SC* | Spectral Clustering |
| *STIP* | Space Time Interest Points |
| *SVM* | Support Vector Machine |
| *SVR* | Support Vector Regression |