AN UNSUPERVISED SEQUENCE-TO-SEQUENCE AUTOENCODER BASED HUMAN ACTION SCORING MODEL

Hiteshi Jain, Gaurav Harit

Indian Institute of Technology Jodhpur

ABSTRACT

Developing a model for the task of assessing quality of human action is a key research area in computer vision. The quality assessment task has been posed as a supervised regression problem, where models have been trained to predict score, given action representation features. However, human proficiency levels can widely vary and so do their scores. Providing all such performance variations and their respective scores is an expensive solution as it requires a domain expert to annotate many videos. The question arises - Can we exploit the variations of the performances from that of expert and map the variations to their respective scores? To this end, we introduce a novel sequence-to-sequence autoencoder-based scoring model which learns the representation from only expert performances and judges an unknown performance based on how well it can be regenerated from the learned model. We evaluated our model in predicting scores of a complex Sun-Salutation action sequence, and demonstrate that our model gives remarkable prediction accuracy compared to the baselines.

***Index Terms*—** Sequence-to-Sequence Autoencoder, Reconstruction, Human action, Scoring

1. INTRODUCTION

Human action assessment has recently gained attention of many researchers in computer vision community, where the objective is to evaluate "*how well*" a person has performed an action.

An automatic human action assessment system has applications in many areas such as sports, health-care, rehabilitation, etc. Assessment of human actions is a difficult task. Such a system needs to learn all possible nuances that an expert trainer has learned over years. The feedback at the end of a performance needs to take into account the entire performance.

The works towards action quality assessment can broadly be divided into three categories : 1) *Action Quality Score Prediction* using regression models[1, 2, 3, 4, 5]; 2) *Human Action Skill Ranking and Classification* into experts and non-experts[6, 7, 8]; 3) *Parametric Assessment*, where only domain specific parameters like grace, consistency, etc. are assessed[9, 10]. As a part of this work we work towards the task of action quality score prediction.

Conventional human action scoring methods[1, 2] used pose features and regressed them against ground truth scores using Support Vector Regression(SVR). Pose features are often wrongly estimated and fail to capture segments of videos that do not involve humans. Recent works [3, 4, 5] instead use 3D convolution features to model human actions and regress these features with the scores. These features outperform pose features, but lack in capabilities as they require a lot of videos to train the system to make it able to predict the scores for a variety of performances.

In order to do the correct scoring, the training data needs to constitute a spectrum of good to bad performances from humans of different proficiency and their respective scores, however, this requires domain experts to annotate large number of action videos and thus is a labor intensive and an expensive task. The question arises : Can we compare the human actions to expert's performance and map the discrepancies to their scores? This would give us an unsupervised technique of human action quality scoring.

In this work we develop a novel unsupervised sequence-to-sequence autoencoder-based assessment model for human action quality score prediction. This model is trained to reconstruct expert performances. Any unseen sequence reconstructed from this trained model would result in generation of a sequence that is interpreted as an adapted benchmark performance which takes into account all the correct performances. We propose a scoring technique where the variations between the input video and the reconstructed video are exploited and the final score for the test performance is evaluated.

The efficacy of the model is tested on *Sun Salutation Assessment Dataset* that we have developed, where the training videos are all expert videos and the test videos have performances of different proficiency levels. The technique is compared with the state-of-the-art regression-based action scoring techniques[1, 3] and template-based assessment technique[10]. It is seen that with fewer number of expert videos and without score annotations, our model outperforms regression models that require wide range of performances and their respective scores.

In section 2, we explain our model for human action quality score prediction. In section 3, we elaborate on our experi-

978-1-7281-2723-1/19/$31.00 ©2019 IEEE

mental setup and finally conclude the paper.

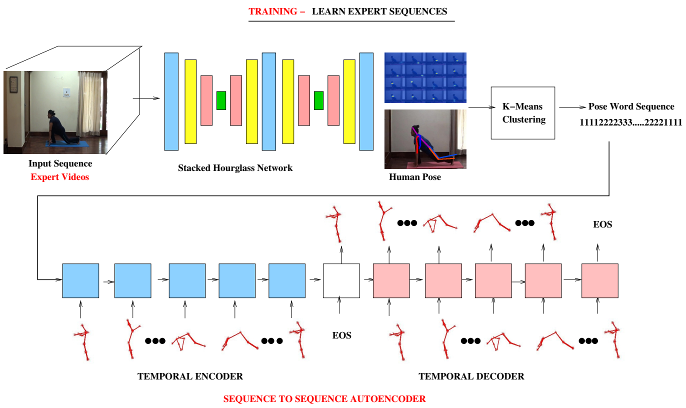

Fig. 1. Sequence-to-Sequence Autoencoder model to learn temporal evolution of expert videos

2. METHODOLOGY

The method described here is based on the principle that as the proficiency of the human performing a certain action decreases, it varies significantly from the expert videos. The variations of a subject's performance from an expert performance leads to penalties that are reflected in the subject's score. We train a sequence-to-sequence autoencoder model that learns the temporal patterns of the human poses across frames. The model is trained with action sequences that consist of expert sequences only, with an objective to minimize the reconstruction error between the input sequence and the output sequence reconstructed from the learned model. After the model is trained, the performances that are close to experts are expected to have low reconstruction error, whereas the sequences consisting of non-experts/amateurs are expected to have high reconstruction error. The reconstruction error can then be used to predict the score of a performer. Our approach consists of three stages : 1) Preprocessing ; 2) Sequence Learning ; 3) Score Prediction

Preprocessing

The task of this stage is to convert raw videos to an admissible input for the model. Following our previous work[10], we use the stacked hourglass networks[11] for human pose estimation. For each frame, the network estimates a pose with 16 joint points (2 for left and right ankles, knee, hip, wrist, elbow and shoulder and for pelvis, neck, thorax, head). The joints of a pose are normalized relative to the head position thus making them translation invariant.

The pose features are reduced to unique pose words (7 in our case) using $K-$means algorithm. This helps us to learn the sequence-to-sequence autoencoder from fewer expert videos. Finally the videos are padded with zeros to give us fixed length videos of size N ($N = 75$ in our case), as an input to the sequence-to-sequence autoencoder.

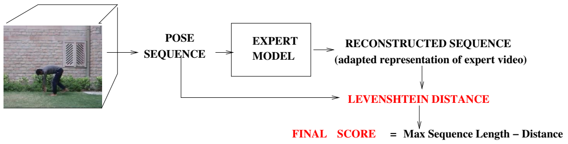
POSE SEQUENCE
EXPERT MODEL
RECONSTRUCTED SEQUENCE
(adapted representation of expert video)
LEVENSHTEIN DISTANCE
FINAL SCORE = Max Sequence Length – Distance

Fig. 2. Scoring of Test Sequences

Sequence Learning Model

Long Short Term Memory architecture[12] can solve many sequence-to-sequence learning problems. We use the sequence-to-sequence learning model as in [13] where the encoder LSTM reads the input pose sequence, one step at a time, and gives a fixed-dimensional vector representation, and decoder LSTM extracts the output pose sequence from that vector (Figure 1). The decoder LSTM is essentially conditioned over the encoder. The LSTM's ability to successfully learn data with long range temporal dependencies makes it a good choice for our application as the score awarded depends on the entire execution sequence.

The goal of the LSTM is to estimate the conditional probability $p(y_1, ..., y_{T'}|x_1, ..., x_T)$ where $(x_1, ..., x_T)$ is the input sequence and $(y_1, ..., y_{T'})$ is its corresponding output sequence whose length may differ from the input length. The LSTM first obtains a fixed-dimensional representation v of the input sequence given by the last hidden state of the LSTM, and then computes the probability of output sequence as :

$$p(y_1, ...y_{T'}|x_1, ...x_T) = \prod_{t=1}^{T'} p(y_t|v, y_1, ..., y_{t-1})$$

Our goal is to develop a model that can well represent all the expert videos. The input and the output of our sequence-to-sequence model are identical. For our work the input sequence is the pose sequence of expert *Sun Salutation* videos. A model trained with the same input and output learns to reconstruct the input video.

We envisage that such a model is able to learn all variations of expert videos. The reconstructed video can be interpreted as a template indicating the correct performance that is most relevant to the input video. This avoids the computations involved in the explicit step of trying out all the templates to choose a right one to compare with as is done in the template based approaches.

Scoring of Test Video Performances

A video performance by a person with high proficiency can be reconstructed correctly using this sequence-to-sequence model trained over all expert videos. However, videos from amateur performers that deviate from these expert videos cannot be reconstructed well as the model has been trained to construct expert videos and the reconstructed video in the case of amateurs would resemble an expert rendering of the action.

The score of a human performance can be calculated using its discrepancy from the expert performance. We use the Levenshtein Distance which gives us the minimum number of single-character edits (i.e. insertions, deletions or substitutions) required to change a pose sequence to its reconstructed output.

In the worst case, when an entirely different action is performed by a subject, the edit distance would be N (the maximum length of the video) and when an expert video is encountered the edit distance would be close to zero. In other words, if the edit distance is denoted by D, the similarity between the reconstructed (expert rendering) and the input pose sequence is given by $N - D$. Figure 2 shows the steps of scoring an test sequence.

This can also be treated as a score of the performer. The range of scores thus would be $0 - N$. To compare the predicted and the ground truth scores, we normalize the scores to a range of $0 - 1$.

In the next section we evaluate our scoring model and compare it with state-of-the-art human action scoring models.

3. EXPERIMENTS

Sun Salutation Assessment Dataset

The assessment datasets proposed in the previous works[1][3] for Diving, Vaults, Figure Skating have a mix of examples of varying proficiency and there are a very few expert videos. Thus to evaluate our idea we collected a new Sun Salutation Assessment Dataset. The dataset is publicly available on our website.

Sun Salutation is a sequence of 7 distinct poses performed in a set cycle forming 10 sub-actions as the human transits gracefully from one pose to another[9].

The training videos of our dataset are organized as two subsets : 1) 35 expert performances to evaluate our model 2) 35 videos which are a mix of expert and non-expert videos to evaluate regression models.

The test set contains 15 videos of varied proficiency, where some videos are similar to experts and others with a variable number of missed sub-actions. The videos were collected from 10 subjects. Figure 3 gives the sample frames from our dataset.

Evaluation Metrics

Baseline and Experiment Settings

We compare our model with 3 baseline works - 1) Pose vs SVR [1], 2) C3D vs SVR, LSTM+SVR [3] 3) Expert Template Matching Approach [10]

For Pose + SVR-based scoring[1], the pose sequences are pre-processed using DCT and DFT operations. We extracted 20 DCT/DFT coefficients from 10 windows of each video to give the final features. For C3D + LSTM-based approach [3], we evaluated the C3D features using C3D model[14] pre-trained over Sports1M Dataset. The LSTM architecture of the scoring network is as proposed in [3].

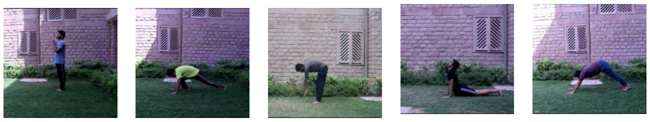

Fig. 3. Dataset sample frames

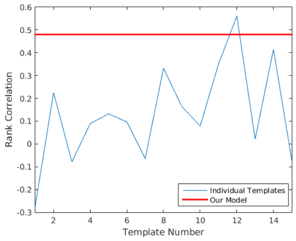

Fig. 4. Rank correlation of individual template videos

For the template based approach[10], and our approach, the poses are converted to 7 codebook words considering 7 distinct poses. The pose-word sequences are used as input to training models. Our architecture has a single layer of LSTM for both encoder and decoder with 64 hidden units for each LSTM layer.

We constrain our baselines to these, as the other models[4, 5] use a segment based approach for scoring, that are not suitable as the videos consist of missed sub-actions and thus the segments in such videos do not cover proper sub-action boundaries.

Similar to [1], we use the Spearman Rank correlation, $\rho = cov(R_p, R_g)/\sigma_{R_p}\sigma_{R_g}$ as our evaluation metric where R_p is the predicted rank by the model (based on predicted scores) and R_g denotes the ground-truth rank of test videos. A higher Spearman correlation implies a better rank prediction. Further, compare the models using the mean square error (MSE) between the $0 - 1$ normalized predicted and the ground truth scores.

Results

Starting with Template Based Approach, we compare the test videos to each of the expert video individually. The Levenshtein distance between the test videos and an expert template is computed to get the scores of all test videos. The rank of the test videos(based on the predicted score) is compared to the ground truth rank to get the Rank Correlation(RC). With experiments it is seen that the rank correlation varies as the expert video changes. Moreover only a single expert template out of 15 templates has rank correlation more than our

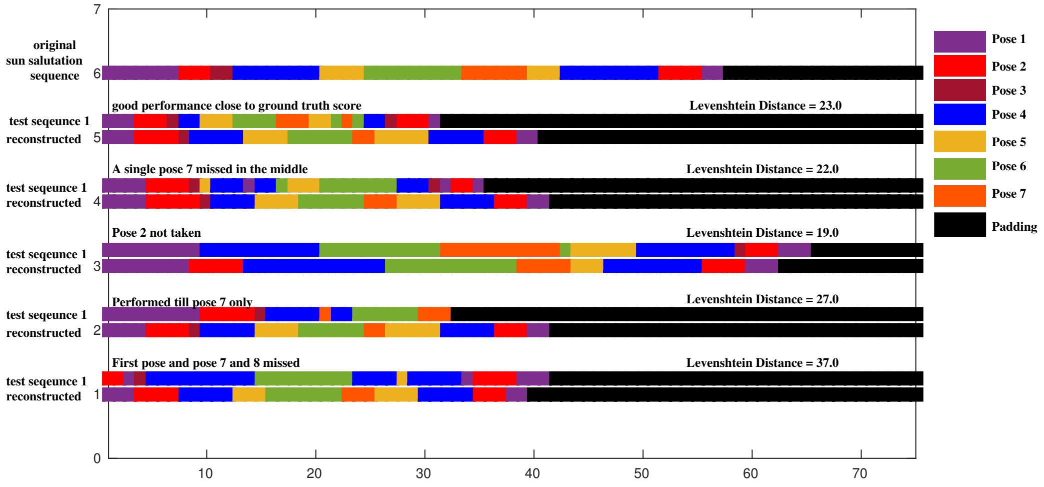

Fig. 5. Reconstruction Results on some sample videos. Top to Bottom: First bar shows the ground truth sun salutation sequence followed by five test sequences and their respective reconstructed videos. Different actions are missed in different videos. The reconstructed sequence matches the ground truth sequence in every case.(Different colors denote different poses. Black color denotes padding used to generate fixed length video)

model(Figure 4). Thus, individual expert videos do not suffice to assess the test performances.

Table 1 gives the comparison results of our model with the supervised regression-based baseline scoring models. It can be seen that with only a set of expert videos, our model outperforms regression based models[1, 3] that requires a mix of correct and incorrect performances. This comes with an added advantage of our model not requiring ground truth scores and thus being completely unsupervised.

Figure 5 shows five sample test sequences and their respective reconstructed sequences. It is seen that the reconstructed sequence is always an adaptation of the ground truth Sun Salutation Sequence (top most sequence on the plot) and serves as the benchmark to compare a given sequence. This is irrespective of the test performance being complete and close to an expert or with variable number of missed poses. (Note : Here we illustrate a single ground truth expert sequence in the figure. However, there can be multiple such expert sequences which may have variable execution speeds.)

It is seen that the variations in speed of the test performance result into variable length reconstructed sequences. Thus the reconstructed sequence is similar to an expert rendering with speed adapted to individual performers.

Thus our model outperforms both the template based matching technique and supervised regression models both in terms of maximum rank correlation and minimum mean square error.

Table 1. Comparison of Rank Correlation and Mean Square Error

Model	**MSE**	**Rank Correlation**
SVR-DCT[1]	0.35	-0.46
SVR-DFT[1]	0.33	-0.39
Pose Words + LSTM	0.18	0.19
Pose Words +LSTM+SVR	0.22	0.23
C3D + SVR[3]	0.23	-0.026
C3D + LSTM + SVR[3]	0.17	0.37
Template Matching[10]	0.33 ± 0.026	0.13
Ours	**0.12**	**0.48**

4. CONCLUSION AND FUTURE SCOPE

We have proposed an unsupervised autoencoder based human action scoring model and evaluated it on complex Sun Salutation assessment dataset. Our model outperforms both the template based and regression models and provides the following advantages: 1) Dataset collection for training the regression models is more tedious because it requires a carefully balanced set of examples in terms of good and bad performances. In contrast our approach requires only expert videos during training. 2) There is no added overhead of annotating the videos with their respective scores during training. The work can be extended to provide feedback to the users and also learn a good ranking model for Olympics actions like diving by including only top rated player videos during training.

5. REFERENCES

[1] Hamed Pirsiavash, Carl Vondrick, and Antonio Torralba, "Assessing the quality of actions," in *European Conference on Computer Vision*. Springer, 2014, pp. 556–571.

[2] Vinay Venkataraman, Ioannis Vlachos, and Pavan K Turaga, "Dynamical regularity for action analysis.," in *BMVC*, 2015, pp. 67–1.

[3] Paritosh Parmar and Brendan Tran Morris, "Learning to score olympic events," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on*. IEEE, 2017, pp. 76–84.

[4] Yongjun Li, Xiujuan Chai, and Xilin Chen, "End-to-end learning for action quality assessment," in *Pacific Rim Conference on Multimedia*. Springer, 2018, pp. 125–134.

[5] Xiang Xiang, Ye Tian, Austin Reiter, Gregory D Hager, and Trac D Tran, "S3d: Stacking segmental p3d for action quality assessment," in *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2018, pp. 928–932.

[6] Aneeq Zia, Yachna Sharma, Vinay Bettadapura, Eric L Sarin, Mark A Clements, and Irfan Essa, "Automated assessment of surgical skills using frequency analysis," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2015, pp. 430–438.

[7] Paritosh Parmar and Brendan Tran Morris, "Measuring the quality of exercises," *arXiv preprint arXiv:1608.09005*, 2016.

[8] Hazel Doughty, Dima Damen, and Walterio Mayol-Cuevas, "Who's better? who's best? pairwise deep ranking for skill determination," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6057–6066.

[9] Hiteshi Jain and Gaurav Harit, "A framework to assess sun salutation videos," in *Proceedings of the Tenth Indian Conference on Computer Vision, Graphics and Image Processing*. ACM, 2016, p. 29.

[10] Hiteshi Jain and Gaurav Harit, "Detecting missed and anomalous action segments using approximate string matching algorithm," in *Computer Vision, Pattern Recognition, Image Processing, and Graphics: 6th National Conference, NCVPRIPG 2017, Mandi, India, December 16-19, 2017, Revised Selected Papers 6*. Springer, 2018, pp. 101–111.

[11] Alejandro Newell, Kaiyu Yang, and Jia Deng, "Stacked hourglass networks for human pose estimation," in *European Conference on Computer Vision*. Springer, 2016, pp. 483–499.

[12] Sepp Hochreiter and Jürgen Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[13] Ilya Sutskever, Oriol Vinyals, and Quoc V Le, "Sequence to sequence learning with neural networks," in *Advances in neural information processing systems*, 2014, pp. 3104–3112.

[14] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4489–4497.