6
Assessment against multiple experts

In the last chapter we discussed an Approximate String Matching approach to compare a test sequence with an expert sequence. The technique was able to provide details of missed and anomalous segments in a test sequence and an overall action assessment. However, it was seen that there are multiple templates possible for an action type which could differ in the count of edits required to match a test sequence. Thus, comparison with a single template is insufficient and we need techniques that can adapt to multiple expert templates while assessing a performance. To this end, we introduce a novel sequence-to-sequence autoencoder-based model which learns the representation using only the expert performances and generates scores for 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 skill assessment accuracy compared to the baselines developed towards human action scoring.

6.1 INTRODUCTION

Template Matching Approaches towards human action assessment differ in feedbacks as the template changes. Among all the possible templates, the correct template for a given test performance is not known aprior. This calls for a technique that can adapt to all the templates while making an assessment. Further, till now we have been talking about reporting missed and anomalous action segments in a performance. Providing an extra incentive in terms of performance scores can encourage people to perform better than their previous execution everytime.

Conventional human action scoring methods [Pirsiavash *et al.* [2014]; Venkataraman *et al.* [2015]] 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 [Parmar and Tran Morris [2017]; Li *et al.* [2018]; Xiang *et al.* [2018]] 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 this 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 chapter, 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.

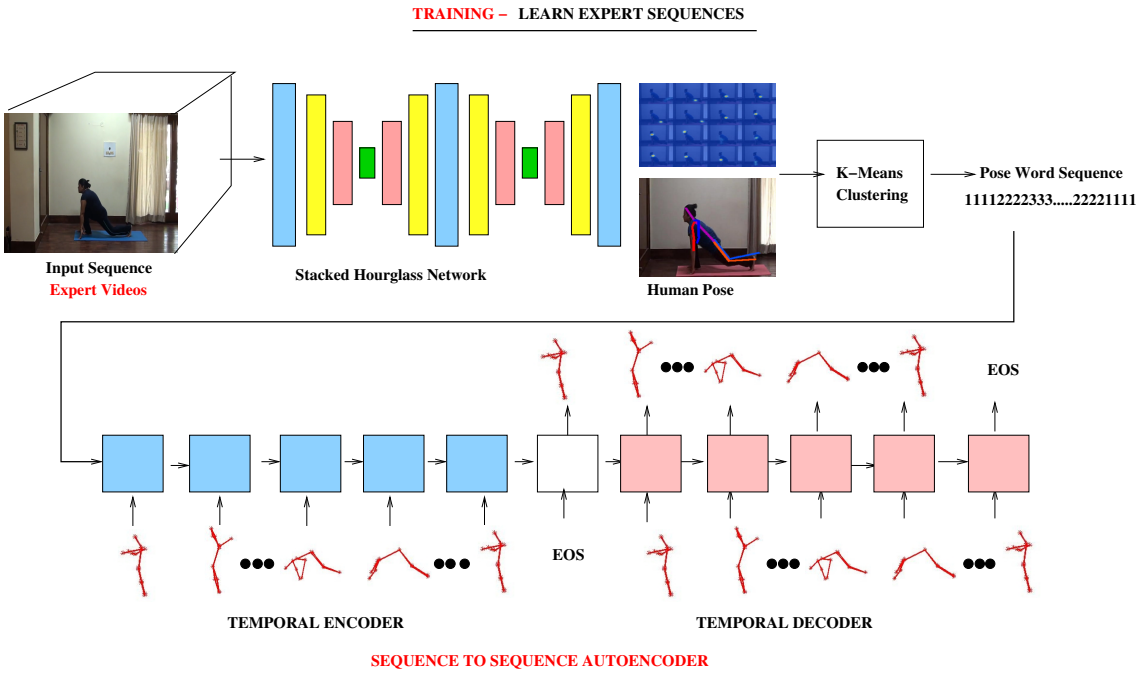

Figure 6.1 : Sequence-to-Sequence Autoencoder model to learn temporal evolution of expert videos

The efficacy of the model is tested on *Sun Salutation Assessment Dataset* that we have discussed in the last section where the test videos have been augmented with their respective scores provided by a yoga trainer. The training videos for our model 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 [Pirsiavash *et al.* [2014]; Parmar and Tran Morris [2017]] and template-based assessment technique proposed in last chapter. 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.

6.2 PROPOSED 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

the same feature representation technique used in the last chapter, we use the stacked hourglass networks [Newell *et al.* [2016]] 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. However in the last chapter we saw that the edits varied even with a little difference in the poses of the performers which arises due to flexibility of performers. With a fixed distance threshold we could not make an allowance of these little variations. Thus here in this chapter we encode pose features to unique pose words (7 in our case) using $K-$means algorithm. This helps us to learn the sequence-to-sequence autoencoder such that it is invariant to little changes in the poses. 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.

Sequence Learning Model

Long Short Term Memory architecture [Hochreiter and Schmidhuber [1997]] can solve many sequence-to-sequence learning problems. We use the sequence-to-sequence learning model as in [Sutskever *et al.* [2014]] 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 6.1). 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

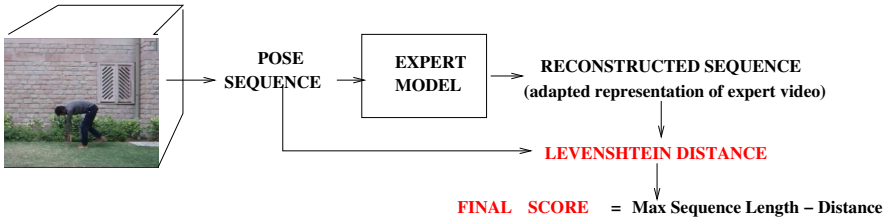

Figure 6.2 : Scoring of Test Sequences

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 6.2 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.

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 6.2 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.

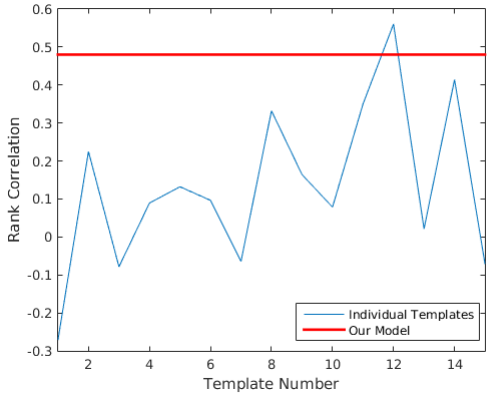

Figure 6.3 : Rank correlation of individual template videos

6.3 EXPERIMENTS

6.3.1 Sun Salutation Assessment Dataset

The assessment datasets proposed in the previous works [Pirsiavash *et al.* [2014]; Parmar and Tran Morris [2017]] 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 updated our Sun Salutation Assessment Dataset as follows :

1. 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.

2. 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.

6.3.2 Baseline and Experiment Settings

We compare our model with 3 baseline works - 1) Pose vs SVR [Pirsiavash *et al.* [2014]], 2) C3D vs SVR, LSTM+SVR [Parmar and Tran Morris [2017]] 3) Expert Template Matching Approach (Discussed in last chapter) For Pose + SVR-based scoring [Pirsiavash *et al.* [2014]], 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 [Parmar and Tran Morris [2017]], we evaluated the C3D features using C3D model [Tran *et al.* [2015]] pre-trained over Sports1M Dataset. The LSTM architecture of the scoring network is as proposed in [Parmar and Tran Morris [2017]].

For the template based approach discussed in last chapter, 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 [Li *et al.* [2018]; Xiang *et al.* [2018]] 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 [Pirsiavash *et al.* [2014]], 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

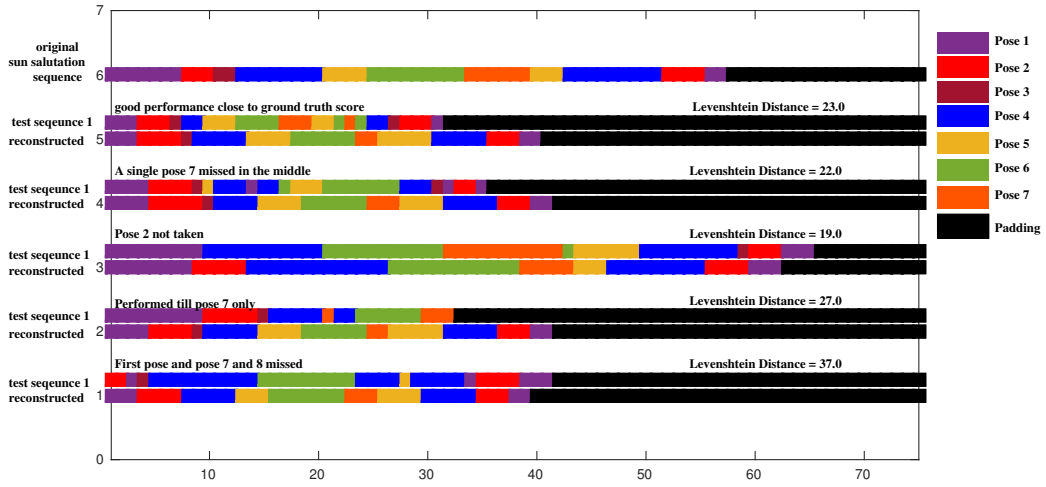

Figure 6.4 : Reconstruction Results on some sample videos. Top to Bottom: First bar shows the ground truth sun salutation sequence followed by five test sequences and their respective reconstructed videos. Different actions are missed in different videos. The reconstructed sequence matches the ground truth sequence in every case.(Different colors denote different poses. Black color denotes padding used to generate fixed length video)

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.

6.3.3 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 model(Figure 6.3). Thus, individual expert videos do not suffice to assess the test performances.

Table 6.1 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 [Pirsiavash *et al.* [2014]; Parmar and Tran Morris [2017]] 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 6.4 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.)

Table 6.1 : Comparison of Rank Correlation and Mean Square Error of various techniques for Sun Salutation performance scoring

Model	MSE	Rank Correlation
SVR-DCT [Pirsiavash et al. [2014]]	0.35	-0.46
SVR-DFT [Pirsiavash et al. [2014]]	0.33	-0.39
Pose Words + LSTM	0.18	0.19
Pose Words +LSTM+SVR	0.22	0.23
C3D + SVR [Parmar and Tran Morris [2017]]	0.23	-0.026
C3D + LSTM + SVR [Parmar and Tran Morris [2017]]	0.17	0.37
Template Matching	0.33 ± 0.026	0.13
Ours	**0.12**	**0.48**

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.

However this approach encodes the poses into fixed number of pose words (7 in case of Sun Salutation). This in turn quantizes the anomalous poses to one of these fixed number of pose words which does not allow us to handle anomalies efficiently. Thus we require a technique that can cluster the poses such that the number of codewords is not fixed and the poses naturally cluster such that the codewords of an anomalous pose varies from one of the key poses.

6.4 CONCLUSION

We have proposed an unsupervised, autoencoder-based human action scoring model that helps compare a test sequence to multiple expert renderings unlike single template matching discussed in the last chapter. Our model outperforms both the template-based and regression models and provides the following advantages: 1) There is no added overhead of annotating the videos with their respective scores during training. Our approach requires only expert videos during training. 2) 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.

However, we saw that encoding the poses to a fixed number of clusters leads to all poses, irrespective of correct or anomalous, to belong to these fixed clusters. Thus the proposed model cannot help identify anomalies. Thus we require a pre-processing step such that the anomalous pose belongs to a cluster different from the key pose i.e. the clustering happens naturally without pre-specifying the count of the clusters.

...