Assessment of Human Actions in Videos

A Thesis submitted by
Hiteshi Jain

in partial fulfillment of the requirements for the award of the degree of
Doctor of Philosophy

॥ त्वं ज्ञानमयो विज्ञानमयोऽसि ॥

Indian Institute of Technology Jodhpur
Department of Computer Science and Engineering
July 2020

9
Conclusion and Future Scope

The objective of the work done in this thesis was to develop methods that can help automate the process of human action assessment. We developed different techniques of how human actions can be compared to reference templates for more objective and interpretable assessment. In this chapter we summarize the main findings with regard to the research objectives and the strengths of the proposed methods are summarized in Section 9.1. The limitations and suggestions for future research are highlighted in Section 9.2.

9.1 SUMMARY OF PROPOSED METHODS AND THEIR RESULTS

Traditional approaches to human action assessment involve human experts who provided a single label action assessment feedback, such as performance score or skill level. A key challenge in action quality assessment is that it is a subjective task with a significant influence of human (expert) bias. The subjectivity in action assessment has been addressed by considering feedbacks given by multiple experts. However, human bias is inevitable. An automated action quality assessment system can bring more interpretability and can be a more objective source of alternate evaluation to subjective evaluation by experts.

The existing works mostly rely on predicting these single labels – skill levels and scores, while optimizing their ranking functions and scoring losses. Though these works have shown promising results in predicting these labels for the test videos, they still lack in their capabilities of providing feedback, i.e. the contributions of sub-actions to the final assessment are not learned well.

In this thesis, we attempt to bring more objectivity to the human action assessment system by considering a set of reference action videos. We have transformed the problem of action assessment into the problem of comparing a given action video with a reference video. This is seen to be a principled approach to deal with expert bias and add interpretability to the action assessment task, as one can assess deviations from the reference video for the temporal segments.

Based on the application domain, we can have a single reference video (eg. rehabilitation), multiple reference videos (eg. physical exercises such as aerobics and yoga), or fewer reference videos (eg. Olympics sports) where there are only few top rated performers. The objective of the thesis is to develop an assessment system for all such scenarios and provide appropriate feedback for test performances.

The research work for achieving the objectives was organized along the following work elements:

i. Comprehensive study of the state-of-the-art methods for human action analysis and assessment

   An understanding of past works in human action analysis is important for building automated action assessment systems. We reviewed the state-of-the-art methods dealing with representation, classification and segmentation of human action videos in Chapter 2. We

discussed the pros and cons of these techniques for different application domains.

Literature on human action assessment has been discussed in Chapter 3. The past work has been categorized based on its general applicability or task specific applicability for human action scoring and skill determination tasks.

ii. We discussed a framework to assess the pace and consistency of a performer. Long term actions like Sun Salutation or Warm Up Exercise are a sequence of postures such that all poses in a cycle should be rightfully taken with smooth transitions between the poses. Further, the consistency in performance needs to be maintained throughout a single cycle and across multiple cycles in case of repetitions. The framework provides feedback on where a wrong transition pace is taken by the performer.

We trained individual Hidden Markov Models over spatio-temporal features for each dynamic posture. The individual HMMs were utilized for automatic segmentation and labeling of the entire Sun Salutation sequence using a concatenated-HMM. The sub-actions of Sun Salutation are not distinctive in nature and have similar motion dynamics for different *asanas* (postures). This often confuses the Viterbi decoder that results into quick changing state sequence. There are techniques to smoothen such quick transitions, however they were seen to give incorrect decoded sequence. We contributed a modified Viterbi decoding algorithm in order to get a smoothened action label sequence. The Modified Viterbi Algorithm resulted into a segmentation accuracy of 81.516% and outperformed the traditional Viterbi that resulted in an accuracy of 58.56%.

The timings for each dynamic posture are compared with the time distributions rendered by multiple experts, to provide feedback for a test subject. The framework has been tested on Sun Salutation Dataset that we developed. The dataset constituted expert sequences and mid-level performers. The proposed framework identifies jerks with a precision of 0.75 and recall of 0.75, while inconsistent actions were identified with a precision of 0.739 and recall of 0.894.

iii. We proposed an exemplar-based Approximate String Matching(ASM) technique for detecting anomalous and missing segments in action sequences performed by amateurs. The technique involves comparing the performed sequence with the benchmark action sequence (as given by experts) and notifying when misalignments occur. This could otherwise not be done using the traditional Dynamic Time Warping template-based matching technique because its boundary conditions lead to alignment failure when there are missing segments.

To evaluate our framework, we extended the Sun Salutation dataset to include amateur performers who missed different intermediate poses in their performances. Our proposed framework notified the missed actions with a precision of 1 and recall of 1 in comparison to DTW algorithm that had a precision of 0.71 and recall of 0.71. Similarly ASM technique notified anomalous actions with a precision of 0.667 and recall of 1 in comparison to DTW algorithm that had a precision of 0 and recall of 0. The precision of our ASM technique in predicting anomalous actions gets affected by posture estimation errors, which led to false alarms. It was observed that DTW could not identify anomalies and gave false alarms due to alignment issues.

iv. Actions like Sun Salutation can be performed in multiple ways. That is, multiple possible variations of the same sequence are equally correct. One such possible variation is the performance speed. The template based approach discussed in Chapter 5 for finding anomalous and missed sub-actions would provide different feedbacks for the same test performance if different expert template are used as reference.

To overcome variations in feedback, we developed a novel unsupervised sequence-to-sequence

autoencoder-based assessment model for human action quality assessment. This model is trained to reconstruct expert performances. For any test performance, the sequences reconstructed by the model are similar to how an expert would have rendered it. Further, the reconstructed sequence also has its speed adapted to that of the test performers and are better indicators of performance quality and do not require selection of the correct template. Deviations between the input video and the reconstructed video are exploited to provide appropriate feedback in terms of score.

The proposed approach was compared to 4 state-of-the-art scoring models. The rank correlation of the predicted scores and the ground truth scores for our approach was 0.48 as compared to the next best C3D+LSTM+SVR scoring model [Parmar and Tran Morris, 2017] which had a rank correlation of 0.37. Additionally, we achieved the lowest mean-square error of 0.12 as compared to C3D+LSTM+SVR scoring model that had a MSE of 0.17. All other baselines gave a poor scoring performance.

The training of a sequence-to-sequence autoencoder-based assessment model requires the pose sequences to be encoded as code-words using techniques like k-means or Gaussian Mixtures. These encoding techniques require the count of distinct poses to be specified before the dictionary words are generated. However, pre-specifying the count of distinct poses is not possible in scenarios when the test sequences can have anomalous poses.

v. We proposed an unsupervised Community Detection-based framework that provides mechanisms to identify key poses in an action sequence without pre-specifying their count. Human actions are composed of distinguishable key poses such that frames around the key poses are mostly similar and thus form dense communities in graph structures similar to friends group on Facebook. This framework helps in identifying anomalous and correct poses as separate communities. This resulted in a better representation of the test videos eventually leading to improved capability of our autoencoder-based assessment framework to provide feedback for test videos. The rank correlation of the autoencoder increased from 0.37 to 0.57 and the mean square error decreased from 0.12 to 0.103

vi. A new approach for action quality assessment using deep learning has been proposed. This approach is applicable in scenarios where too few training videos are available for any particular type of action such as Olympics. An LSTM-based Siamese network is used to learn discriminative features from pairs of videos characterized as similar and dissimilar using the groundtruth scores. The learned model is then used for action scoring, where the performances are compared with the reference expert performance to determine the score. This enables interpretability to the given score, as a comparison of temporal segments with the reference video segments is possible with the learned model.

We tested our scoring model to 3 baseline works for two actions $-$ Diving and Gymnastic vaults from UNLV dataset [Parmar and Tran Morris, 2017]Parmar and Morris [2019]. For the diving action, the rank correlation of the predicted scores and the ground truth scores for our approach was 0.69 as compared to the next best C3D+LSTM(final label) scoring model [Parmar and Tran Morris, 2017] which had a rank correlation of 0.64. Additionally, we achieved the lowest mean-square error of 88.59 as compared to C3D+LSTM+SVR scoring model that had a MSE of 94.97. All other baselines gave a poor scoring performance. For vaults our proposed model achieved a rank correlation of 0.53 as compared to C3D+LSTM scoring model that achieved an RC of 0.34. In contrast the proposed model achieved a lower means square error of 0.36 compared to C3D+LSTM that achieved a mean square error of 0.69.

Further, the Siamese network trained to learn the similarity metrics for a complete action was used to find sub-action level similarities. It was seen that the network was able to highlight the

discrepancies in the performed action relative to expert videos.

9.2 FUTURE SCOPE

Human action assessment is a difficult task. An automated system for action quality assessment needs to learn all possible nuances that a human expert learns over years. Numerous research facets emerging out of the works presented in thesis may provide worthwhile exploration directions to researchers for their endeavors. Some of the main directions can be identified as follows:

i. Generation of Action Assessment Captions: Providing human action assessment feedback as captions is mostly unexplored. This may be attributed to the lack of reliable ground truth for the assessments as it is a subjective task. However, it would be interesting to explore this direction.

ii. Most of the action scoring works have been tested for the single view action like diving. It is seen that the performance of the action quality assessment techniques deteriorates as actions with view variations are included. Thus, designing assessment systems that can adapt to view variations in action classes can be a good contribution to the domain.

iii. Multi-modal Assessment: Events like Olympics involve audience reactions too, like cheers when there are good performances, gasps during bad performances. It would be interesting to include the audio features along with the visual performance features to assess such events.

iv. Cognitive perspective can be included during model training. This will allow training models using cognition data from non-expert observers of performance.

...