3
State-of-the-Art : Human Action Assessment

In the last chapter, we discussed the strength and weaknesses of various human motion representation and activity modeling techniques to meet the requirements of automated monitoring systems in various domains. The recent years has seen many such systems introduced for the end users. These automated, intelligent, vision based monitoring system can assist users in various tasks and can aid them in the process of risk detection and analysis. In this chapter we review some of the past works towards automated vision-based human action assessment.

The task of human action assessment aims at monitoring the perfection with which an action has been performed. There can be various important parameters that define what is a perfect action. The best action is the one which results in the best reward for the performer. The reward can be subjective and not always objectively realised by an outside viewer, such as, improvement in performer's concentration from yoga. Automated human action assessments can only address objective targets such as score evaluation, skill assessment, correct posture evaluation.

Early works towards human action assessment were hand-crafted for certain applications. Though these methods performed well they could not generalize to scenarios with different action types. There is a need for generic learning-based human action scoring model that could be generalized to all action types. The task of skill assessment is another direction of work that has been introduced recently where the target is to evaluate if one performance is better than the other or vice-versa. We discuss these three directions in detail and discuss the motion representation techniques for human action assessment.

3.1 MOTION REPRESENTATION FOR HUMAN ACTION ASSESSMENT

From the perspective of motion representation the reviewed methods can be divided into three categories:

1. Skeleton-based representation - In order to develop an accurate action evaluation method, it is preferred that the dynamic changes in joints or body parts are used for action analysis. In the early works, skeleton data was mainly captured from Kinect cameras [Su [2013], Zhao *et al.* [2014]]. However, depth cameras severely suffer from occlusions, sensing distance, and poor performance in outdoor environments. With significant progress in recent pose estimation techniques and methodologies, skeleton data can now be estimated from RGB image data. Traditional skeletonization models - flexible mixtures of parts model [Yang and Ramanan [2011]] have been used by Pirsiavash *et al.* [2014]. The estimated skeleton data can often be noisy in cases of occlusion or cluttered backgrounds. The estimated skeleton data can often be noisy in cases of occlusion or cluttered backgrounds in in realistic scenes. To obtain robust similarity quantization for fine-grained quality assessment, the preliminary treatment of skeleton data is noise filtering. The traditional filtering techniques like DCT and DFT have been employed to transform the discrete coordinates of the joint trajectory. Consequently, zero values or sharply changing coordinates can be filtered out, and low-frequency components are preserved for the reliable detection of human body positions. Since both the height of the persons and the camera

distance are quite different, the scale of the human body in different videos can be quite diverse. Thus the original skeleton positions should be normalised for comparisons. This is done using scale normalisation process where the head location is subtracted from all the joint locations. One of the drawbacks of using human poses as a feature is that the scoring is affected by incorrectly estimated poses. Moreover, pose-based representations are incapable of modeling objects used during an action (such as sports ball or tools), and these do not consider physical outcomes (such as splashes in diving) that may be important features for some activities. However, these features are important while providing an interpretable feedback on how the performer can improve his action.

2. Handcrafted spatio-temporal features - These features have been extensively used for human action recognition task but have not been used in practice for assessment works. Spatial–temporal interest points (STIP) detector [Laptev [2005a]] has been used in [Zia *et al.* [2015]Zia *et al.* [2016]] for classifying the performers according to their surgical skills.

3. Deep learning-based spatio-temporal features - Deep features such as C3D, P3D features learned from 3D convolutional neural networks have recently been utilized to represent human actions during the task of human action scoring [Parmar and Tran Morris [2017]; Xiang *et al.* [2018]; Li *et al.* [2018]]. The feature extraction has been done in two ways - a) using pre-trained models [Parmar and Tran Morris [2017]; Li *et al.* [2018]] b) using an end-to-end setup for learning the features with assessment objective (scoring in case of AQA and ranking objective in case of skill assessment) [Xiang *et al.* [2018]; Doughty *et al.* [2018a]]. The extracted features are aggregated using LSTMs, average pooling or concatenation techniques.

3.2 HAND-CRAFTED ASSESSMENT SYSTEMS

We make a domain wise discussion of works when we talk about hand-crafted assessment systems. We particularly talk about healthcare, entertainment, sports and exercises and actions of daily living.

Heathcare

Automatic monitoring systems have contributed healthcare domain in two ways - (i) by providing real-time rehabilitation platforms, and (ii) the skill assessment platforms for the doctors themselves.

The systems designed to provide rehabilitation platforms to the user compare the input postures of the people with a predefined valid template. The user is warned as soon as an incorrect posture is seen and feedbacks are given to improve. These systems are mostly based on human pose acquired from the Kinect sensor.

A Kinect sensor is available commercially for around $100, which is about the cost of a single physical therapy session. These cameras are easy to install at home and facilitate proper performance of rehabilitation exercises, which would lessen visits to a physical therapy center. Thus, these monitoring systems prove to be profitable to the end user.

Su [2013] proposed offline estimation of feedback for Kinect-based rehabilitation exercises by comparing the performed motion of the person with his own execution recorded previously in the presence of a professional. The comparison was done using dynamic time warping (DTW) with Euclidean distance as similarity measure. The system provided qualitative feedback (eg. your arm positions are too high) as fuzzy logic on the similarity of body joints and execution speed but does not suggest the user how to correct the movement. Further, the framework used is computationally

intensive and hence the assessment is done off-line and not in real-time.

Zhao *et al.* [2014] described a rule-based system that facilitates in-home monitoring and guidance for rehabilitation exercises. For exercises like hip abduction and bowling, a set of correctness rules were defined which are independent of the size of the users. The performance of the person was represented in terms of Kinect joint points and angles which were compared with the saved rules to assess the performance of the person. The system gives quantitative feedback on which rules were violated.

Further, some kinect-based real time physiotherapy platform [Pachoulakis *et al.* [2016]] were reported recently tailored to Parkinson's disease (PD) patients. A small collection of exercises practiced in traditional physiotherapy for PD patients has been implemented in the Unity 3D game engine. An individual patient's performance is compared and performance feedback is provided to the patient.

In [Zia *et al.* [2015] and Zia *et al.* [2016]], an automated framework was proposed for visual assessment of the expertise level of surgeons. The surgical tasks are represented as a sequential stream of Space Time Interest Point features transformed into frequency domain components. The frequency domain transformation effectively extracts the information that can differentiate the different skill levels of the surgeons. The basic surgical tasks such as suturing and knot tying are sequential and periodic activities. An expert surgeon can perform these tasks more number of times in a given time as compared to a beginner. Thus in terms of signal, the expert 'signal' will have a higher frequency as compared to a beginner or intermediate. The assessment technique exploits the periodicity of the two surgical tasks and distinguishes the three skill levels of the surgeons - expert, intermediate, and beginners.

Sports and Exercises

Gordon [1995] used tracking algorithms to assess gymnastic vaults. The authors recorded the vault performances using Hi8 video recorder camera. The frames of the videos were then analyzed using a motion tracking algorithm that provided the centre of a moving object in a sequence of frames. The position data computed using tracking algorithms were then used to calculate the deductions in score according to FIG (Federation Internationale de Gymnastique) Code of Points that pertain to the location of the performer while he performs gymnastic vaults. This includes rules for incorrect flight-path trajectory and insufficient distance and height.

Team sports like basketball are defined by a group of people who tend to organize in groups to achieve certain goals. Jug *et al.* [2003] developed a probabilistic model of a basketball team play, based on key events detected in the team behavior. The model is used to assess the team performance in different types of basketball offenses, based on trajectories of all players, obtained by whole-body tracker.

Parisi *et al.* [2016] proposed a novel recursive neural network that uses growing self-organization for the efficient learning of body motion sequences from power-lifting exercises. The quality of actions is then computed in terms of how much a performed movement matches the correct continuation of a learned sequence. The proposed system provides visual assistance to the person performing an exercise by displaying real-time feedback.

Entertainment

Any of the existing dance forms can be learned from two main resources: (1) learning from a dance instructor, or (2) self-learning from videos. The expertise of the instructor, with their ability to give real-time feedback on how well the student performs makes it an obvious choice. However, the expenses of most choreographers or their unavailability at all times make it difficult for the user to

choose this option. Self learning from videos is not a good choice either because there is no feedback for a wrongly performed step.

The advances in vision-based monitoring systems have brought in a new way to learn the dance forms using visual dance monitoring system where the student learns from the posture hints provided by the system from pre-saved dance posture sequence and the postures taken by the student are analysed by the same system to provide real-time feedback. We discuss some of them here.

Alexiadis *et al.* [2011] and Alexiadis and Daras [2014] proposed a system for automatic dance performance evaluation against a standard performance is from human Motion Capture (MoCap) data. The framework addresses global and local temporal synchronization, spatial alignment and comparison of two "dance motion signals." A quaternionic representation over the 4D poses is used and a global temporal synchronization of dance MoCap data is achieved against the teacher's performance by the use of quaternionic cross-correlations. A set of quaternionic correlation-based measures (scores) are proposed for evaluating and ranking the performance of a dancer. The performance of the system is analysed over Huawei/3DLife/EMC2 dataset.

Kyan *et al.* [2015] proposed a framework to assess and visualize ballet dance movements performed by the students in an instructional, virtual reality setting. The framework is based on the unsupervised parsing of ballet dance movement into a structured posture space using the spherical self-organizing map (SSOM). A unique feature descriptor is given to the different ballet dance movements, which are gesture trajectories through posture space on the SSOM. This allows the system to identify the category of movement the student is attempting. The dance sequence is compared against a library of gestural components performed by the teacher. A virtual feedback system is also presented.

Activities of Daily Living

Wu *et al.* [2015b] modeled human activities comprising multiple actions in an unsupervised setting. They learned high-level action co-occurrence and temporal relations between the actions in an activity video. The learned model is then applied to a novel application that detects forgotten actions using action patching. The query video is segmented into action classes and the missed actions from the set of learned action classes are patched at each segmentation point in the segmented sequence. Its probability of occurrence at the segmentation point is then evaluated. A high probability signifies that the action is missed at that point in time.

[Soran *et al.*, 2015] proposed a novel system that understands the ongoing action before it completes and decides whether there is a missing action or not for generating notifications about this missing action. For this the action is modeled to exist in one of the three states : action beginning, middle or end. The on-going action can be thus decided just by the initial few frames of the action, using an HMM-based action part classifier. The inter-action dependencies are represented using flexible ordered graph for which edge weights are set inversely proportional to the probability of seeing two actions together in the set of training videos. The missing actions are detected if the user takes a path which is not the shortest path on the graph.

[Costa *et al.*, 2016] attempted to develop automatic meal intake monitoring system that helps in tracking people's eating behavior. Kinect based features are analyzed to detect eating gestures which are then classified using hidden markov model (HMM) to evaluate the eating behavior of the elderly people.

Another application field is the assembly line operations in industrial production. Human action assessment can help to construct standardized action models related to different operation steps, as well as to evaluate the performance of trained workers. Automatically assessing the action quality of

workers can be beneficial by improving working performance, promoting productive efficiency, and, more importantly, discovering dangerous actions before damage occurs.

Till now we have been discussing domain specific assessment techniques that have been introduced in past. These techniques do not generalize to all action types. For example, the trajectory based gymnastic vaults assessment cannot be exactly replicated to diving as the rules differ for both. We discuss more generic action assessment works next. These works are divided into two broad categories: Human quality action assessment and skill determination. We discuss both of these categories in detail.

3.3 HUMAN ACTION QUALITY ASSESSMENT (AQA)

Human action quality assessment (AQA) or action scoring is a recent researched topic in the computer vision community. It aims at developing generic scoring models for all action types. Scoring is posed as a supervised regression task, where the human action features are regressed against the ground-truth scores annotated by a domain expert and the learned model is used to predict the scores for a test performance sequence.

We discuss the past human action scoring models under three: 1) regression framework and score / rank loss functions 2) Feedback Proposals 3) Action Scoring Datasets

3.3.1 Regression Models

Scoring models introduced in the past works come in two variations : 1) Linear Regression and 2) Support Vector Regression. Of special consideration, is the fact that the scoring models by Pirsiavash *et al.* [2014], Venkataraman *et al.* [2015] and Parmar and Tran Morris [2017] optimize the mean square error between the ground truth and the predicted score. While Li *et al.* [2018] introduces an additional rank loss with the mean square loss to optimize the network. Mathematically, for a batch with n samples, s_i denoting the predicted score and g_i denoting the ground truth score

$$L_1 = \frac{1}{2n}\sum_{i=1}^{n}(s_i - g_i)^2$$

describes the means square error loss function common to all works, and

$$L_2 = \sum_{i=1}^{n}\sum_{j=1,j>i}^{n} RELU(-(s_j - s_i)sign(g_j - g_i))$$

denotes the ranking loss function introduced in Li *et al.* [2018] and is used in addition to the mean square error loss. The ranking loss ensures the right order of predictions. When the predictions violate the ranking constraint, the ranking loss will generate a punishment term. Otherwise, the value of the ranking loss is zero.

All these approaches work towards absolute score estimation. However, a common practice of judgement is to score a performance relative to an expert execution. We include the concept of relative scoring in our scoring framework and discuss it in Chapter 8.

3.3.2 Automated Feedback Techniques

The works discussed so far introduced techniques towards providing feedback at sub-action level based on the final score. Pirsiavash *et al.* [2014] provide feedback on how the performer can improve his/her action. They provide directions where the body parts should move to maximize the score. This is accomplished by differentiating the scoring function with respect to the joint location. Maximizing the gradient of the score with respect to the location of each joint can find the joint direction that the performer must move to achieve the largest improvement in the score.

Parmar and Tran Morris [2017] introduced a new training protocol called *incremental label training* to provide sub-action level scores. It is expected that as an action advances in time, the score should build up (if the quality is good enough) or be penalized (if the quality is sub par). Following this intuition, they believed that the score should be accumulated throughout an action as a non-decreasing function. Thus instead of using just the final score label for the whole action instance, they provided an intermediate score label which was the score accumulated till a particular clip. A video is divided into equal size non-overlapping clips with an assumption that each clip contributes the same amount of score. Incremental-label training is used to guide the LSTM during the training phase to generate the final score along with intermediate score outputs (i.e. back-propagation occurs after each clip).

The temporal score evolution as it changes through the LSTM structure is utilized to identify both "good" and "poor" segments of an action. The assumption is that a perfectly executed action would have a non-decreasing accumulation of score while errors will result in a loss of score.

Both these feedback techniques are prone to errors: The pose-based feedback introduced in [Pirsiavash *et al.* [2014]] is the most interpretable form of feedback, however, its performance is highly dependent on the pose estimation accuracy. The incremental learning is based on an assumption that the scores are evenly distributed in sub-segments which is not realistic too.

3.3.3 Olympics Action Scoring Datasets

The past works have contributed Olympics dataset for 3 actions : Diving, Gymnastic Vaults and Figure Skating. In our work we consider the first two since we observed that Figure Skating is too random an action for automatic judgements.

Diving Dataset: The MIT-Diving dataset [Pirsiavash *et al.* [2014]] consists of 159 videos taken from 2012 Olympic men's 10-meter platform prelims round. The UNLV-diving dataset [Parmar and Tran Morris [2017]] is an extension to this dataset, which includes semi-final and final round videos, totaling to 370 videos, each having roughly 150 frames. The scores vary between 0 (the worst) and 100 (the best). A diving score is determined by the product of execution score (judging quality of a diving) multiplied by the diving difficulty score (fixing agreed-upon value based on diving type).

More recently Parmar and Morris [2019] proposed a Multitask AQA Dataset with 1412 samples of diving videos including 10m Platform as well as 3m Springboard with both male and female athletes, individual or pairs of synchronized divers, and different views. The AQA score, dive type and the commentary for each sample was included in the dataset.

UNLV-Vault DatasetParmar and Tran Morris [2017]: This dataset includes 176 videos. These videos are relatively short with an average length of about 75 frames. The scores range from 0 (the worst) to 20 (the Best). A vault score is determined by the sum of the execution score and the difficulty score.

3.4 HUMAN ACTION SKILL DETERMINATION

While human action scoring is aimed at predicting scores of performers, the literature on skill determination identifies relation between two performers and determines if a video is better than the other or similar in skill. In Doughty *et al.* [2018a], the problem of skill determination is addressed for tasks like surgery and rolling pizza dough. The model learns shared features when two videos in a pair are equivalent in terms of skill level and learn discriminative features when they exhibit variance in the skill levels. The videos are uniformly divided into three clips. Clips from two videos are fed in pairs into a Siamese network with shared weights. It is assumed that the overall skill is maintained across all clips, i.e. if video i is a better skill video than j, then all the three segments of video i would inherently be better than that of video j. The network is then trained to learn a feature representation f such that the function can define the skill level of the videos. This is learned using $0-1$ ranking error loss:

$$L_{rank} = \sum_{(p_i, p_j \in P)} max(0, m - f(p_i) + f(p_j))$$

where *P* is the set of all pairs of training videos.

In addition, an adversarial loss is used to distinguish between similar pairs of videos. For similar videos, $f(p_i) \approx f(p_j)$, thus the rank loss described above changes to :

$$L_{sim} = \sum_{(p_i, p_j \in P)} max(0, |f(p_i) + f(p_j)| - m)$$

These two losses are jointly optimized to learn both video pairs where skills are similar and the video pairs where the skills are different. A variation of this approach was proposed by Doughty *et al.* [2018b] where instead of assuming an equal contribution of all clips in skill determination, a temporal attention model is proposed, alongside ranking model, such that the clips are now seen to provide unequal contribution to the final ranking.

3.5 OPEN CHALLENGES

In this chapter, we discussed some of the works towards human action assessment. The existing literature was divided into three categories: i) Action specific assessment, ii) Action Quality Assessment or Action Scoring iii) Skill determination. The works proposed in the last two categories are generalize well to many action types and reported good assessment performance. However, there are some open challenged that still exist.

The Action Quality Assessment and Skill Determination works depend on a single label to train the square error loss and ranking loss and develop models to predict scores and skill respectively. These labels are provided by the expert judges and lack interpretation. For e.g. a score provided for a full Olympic diving does not explain where the score got deducted.

Action assessment is a subjective task with a significant influence of human judgement bias, thus leading to less reliability. The biases may include a nationalistic bias [Ansorge and Scheer [1988]; Emerson et al. [2009]], where judges give higher scores to performers from their own countries, a difficulty bias [Morgan and Rotthoff [2014]], where athletes attempting more difficult routines receive higher execution scores, etc. Traditionally, the reliability issue in action assessment has been addressed by considering feedbacks given by multiple experts. Nevertheless, this solution is less affordable and does not entirely get away with subjective bias.

In a recent study [Mazurova and Penttinen [2020]], the authors collaborated with Finnish Gymnastics Association and conducted interviews with gymnasts, coaches and directors to understand their perception on electronic judging systems. The study revealed a positive side of the transition to electronic judging systems that mainly relates to the deficiencies of the human-based judging, it being vulnerable to biases, human error , human fatigue, judges' personal preferences, and inherent lack of explanation. The informants expressed that electronic judging systems contain affordances that could efficiently mitigate the said challenges associated with human-based judging. Thus, it will be interesting to develop an automated action assessment system bring more interpretability and little objectivity to this domain.

We believe that a principled approach to deal with expert bias and to provide an interpretable feedback is to consider the problem of action assessment can be transformed into the problem of comparing a given action video with a reference video which is the top-rated performance video.

The simplest method to compare a test video with expert execution is to use Dynamic Time Warping as a template-based matching technique. However, the Dynamic Time Warping technique is not designed to handle missing data or anomalous sub-segments due to its boundary constraints and monotonicity conditions, thus leading to wrong alignments of the two video sequences. Thus there is a requirement of approaches that can help compare the performances with the reference videos and provide a more detailed interpretation.

In this thesis, we address the problem of assessment as a problem of comparing a given action with reference videos under three scenarios: 1) single reference video-based assessment, 2) reference collection with many expert templates, and 3) reference collection with fewer experts. The proposed assessment techniques have been evaluated for two application scenarios - Yoga and Olympics events. We are able to collect multiple expert sequences for Yoga while Olympics MIT and UNLV datasets [Pirsiavash *et al.* [2014]; Parmar and Tran Morris [2017]] constitute few top scores only. The developed methods are applicable to diverse action types.

We discuss our key contributions towards human action assessment task in the following chapters.

...