1
Introduction

Understanding human actions is a very important research problem in computer vision domain. Traditionally, it has been primarily explored from the perspective of action recognition, i.e., identifying the type of action being performed in the video irrespective of the variations in appearance, view, self-occlusions, and background clutter. The emerging popularity of the area of human action analysis can be attributed to the important role of such techniques in commercial vision-based monitoring systems.

Automatic monitoring systems have become an integral part of different application domains: healthcare [Zia *et al.* [2015]], sports and exercises [Pirsiavash *et al.* [2014]], entertainment [Alexiadis and Daras [2014]], activity of daily living [Wu *et al.* [2015a]], etc. For example an automated scoring system for Olympics events like Diving and Gymnastic Vaults could be used as an alternate opinion to avoid judgment biases.

Action recognition involves discriminating between two very different classes, while action assessment requires identifying differences in the quality of action videos within a class, thus making the assessment task more difficult.

The first concern while considering action assessment is to determine whether an action is appropriate for assessment using computer analysis. In other words, is it possible to develop an evaluation metric that can adequately provide an evaluation just by using the visual information? For many actions it is not always easy to formulate a metric to evaluate the goodness of a performance. This is especially true for artistic and unconstrained performances such as dance or gymnastic skating. Experts judge the quality of such performances based on an intuitive sense, and it is difficult to explicate their evaluation criteria into formal rules or trained models due to creative moves taken by performers.

The scope of this thesis is limited to development of assessment techniques for physical actions that are executed in a fixed manner. Examples of such actions include gymnastic vaults, diving, aerobics, yoga, etc. We discuss the research gaps in the existing literature in human action assessment, which motivated us to look upon this area of research. Section 1.4 enlists our contributions and Section 1.5 presents the organisation of chapters in the thesis.

The field of human action assessment is relatively new, and there are only few works that have been introduced in this domain. Existing works towards human action assessment can broadly be divided into two categories: 1) Action Quality Assessment [Pirsiavash *et al.* [2014]; Venkataraman *et al.* [2015]; Parmar and Tran Morris [2017]; Xiang *et al.* [2018]] where the goal is to predict the scores of the performances in a video depicting action types like Olympics diving and gymnastic vaults. These works regressed video representations to their respective scores. 2) Skill Assessment that evaluates the skill of performers using a pairwise deep ranking model which learns to rank a video relative to another based on the difference in skill [Doughty *et al.* [2018a,b]; Fawaz *et al.* [2018]; Zia *et al.* [2015]]. Both categories use a single target - final score or skill-level to train the regression or ranking models. 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.

1.1 CHALLENGES / PROBLEMS ADDRESSED IN THESIS

The objective of this thesis is to develop machine learning based approaches for assessment of human actions in video. The following problems have been addressed:

1. Many times an action being performed may have missing steps or incorrectly performed steps (anomalous sub-actions). Detection of missing or anomalous portions of an action video is a challenge for existing techniques that attempt to do a temporal alignment of a given performance with a benchmark template.

2. Many actions can be performed correctly in multiple ways with flexibility in terms of speed variations. A single model has difficulty in learning these multiple templates and allowable speed variations.

3. Assessment intrinsically involves *comparison* with the 'correct' way or the 'expert' way. This poses a generic problem of how actions can be compared? A system that can learn to compare action sequences is more generalizable to different types of actions. Such a network can be used to assess/score smaller segments of an action and thus give more interpretable assessment.

1.2 PROBLEM FORMULATION

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.

1.3 MOTIVATION FOR THE APPROACHES DEVELOPED

Action assessment is a subjective task with a significant influence of human judgement bias, thus leading to less reliability. 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. Thus, it will be interesting to develop an automated action assessment system that can bring more interpretability and objectivity to this domain.

A principled approach to deal with expert bias and to add more objectivity to the assessment is to consider a set of reference action videos where the action was performed with high precision (based on expert scoring). Thus, the problem of action assessment can be transformed into the problem of comparing a given action video with a reference video.

The simplest method to compare a test video with expert execution is to use Dynamic Time Warping as a template-based matching technique. It has been used to design a personal rehabilitation exercise assessment to determine the similarity between the patient's exercise and a standard template Su [2013]; Palma *et al.* [2016]. However, amateurs or first-time performers, tend to forget action steps and perform some unwanted action movements while performing a long term action sequence. The traditional 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. We provide solutions to identify such segments in a video.

Further, in practice, there can be many executions of the same actions that can be considered as ideal templates. These templates can vary due to variations in speed and flexibility of performers. Template matching approaches towards human action assessment fails as among all the possible templates, the correct template for a given test performance is not known aprior. Thus a technique that can adapt to all the templates while making an assessment is essential.

The count of ideal templates or the reference action videos can vary with different application domains. In applications like rehabilitation or yoga we can get many ideal executions from expert trainers. In contrast, the top rated performers in Olympic events like diving and gymnastic vaults are too few. Thus we need assessment models for both scenarios.

1.4 OUR CONTRIBUTIONS

The contributions of this thesis are as follows:

1. A new Sun Salutation (*Surya Namaskar*, a famous Yoga practice) dataset has been constituted, which consists of performers of different skill levels - expert, intermediate, and amateur performers with their respective scores and feedbacks.

2. A framework to assess the Sun Salutation performance on the parameters of grace and consistency has been developed. The individual inter-pose timings are assessed against the experts' timings. The long-term actions are segmented using Hidden Markov Model. A new Modified Viterbi Decoding algorithm has been proposed to handle cyclic actions like Sun Salutation.

3. An Approximate String Matching-based single template matching approach has been proposed to identify missed and anomalous segments in the Sun Salutation performance.

4. A sequence-to-sequence Autoencoder model has been proposed to assess the performed Sun Salutation sequence against its multiple expert executions. The autoencoder model learns a unified representation using the expert performances and judges an unknown performance based on how well it can be regenerated from the learned model.

5. The input representation in the form of pose sequence to the autoencoder has been improved using an unsupervised community detection-based technique to cluster frames (poses). The proposed approach overcomes the limitation of traditional clustering algorithms which require the number of clusters k to be pre-specified and then incorrectly group the anomalous poses to be clustered to one of the k key poses.

6. 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. An LSTM-based Siamese network is used to learn discriminative features from pairs of videos with similar and dissimilar 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 temporal comparison with the reference video is possible with the learned model.

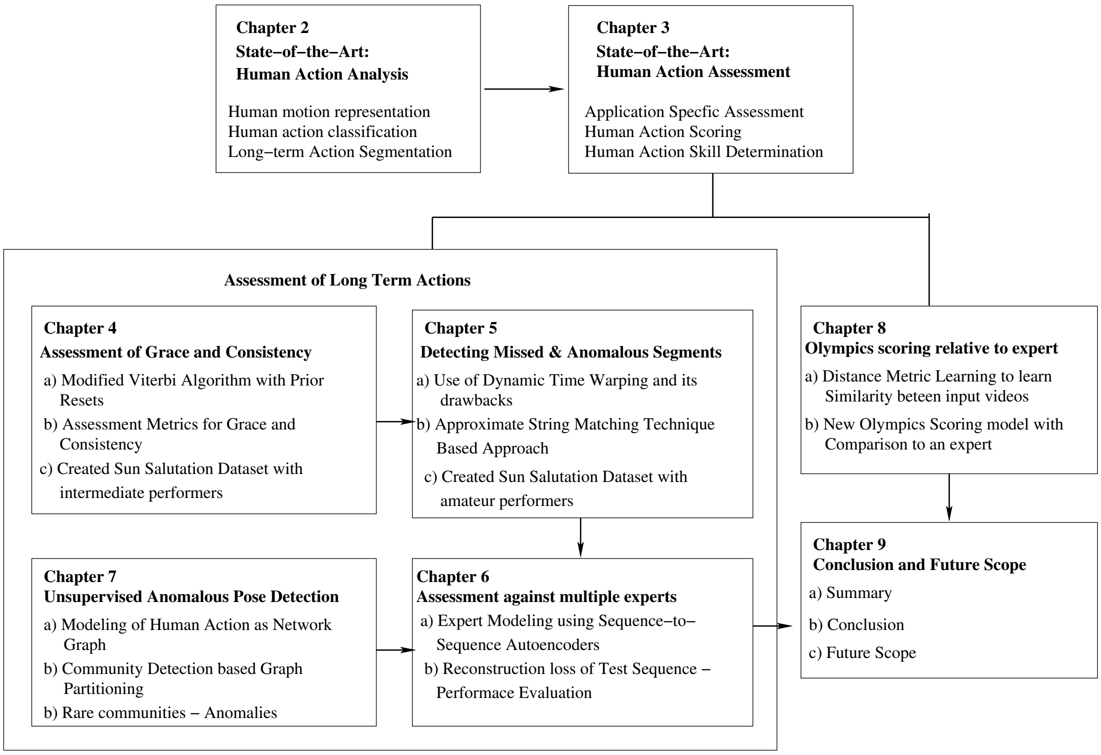
Chapter 2
State–of–the–Art:
Human Action Analysis
Human motion representation
Human action classification
Long–term Action Segmentation
Chapter 3
State–of–the–Art:
Human Action Assessment
Application Specfic Assessment
Human Action Scoring
Human Action Skill Determination
Assessment of Long Term Actions
Chapter 4
Assessment of Grace and Consistency
a) Modified Viterbi Algorithm with Prior Resets
b) Assessment Metrics for Grace and Consistency
c) Created Sun Salutation Dataset with intermediate performers
Chapter 5
Detecting Missed & Anomalous Segments
a) Use of Dynamic Time Warping and its drawbacks
b) Approximate String Matching Technique Based Approach
c) Created Sun Salutation Dataset with amateur performers
Chapter 8
Olympics scoring relative to expert
a) Distance Metric Learning to learn Similarity beteen input videos
b) New Olympics Scoring model with Comparison to an expert
Chapter 9
Conclusion and Future Scope
a) Summary
b) Conclusion
c) Future Scope
Chapter 7
Unsupervised Anomalous Pose Detection
a) Modeling of Human Action as Network Graph
b) Community Detection based Graph Partitioning
b) Rare communities – Anomalies
Chapter 6
Assessment against multiple experts
a) Expert Modeling using Sequence–to–Sequence Autoencoders
b) Reconstruction loss of Test Sequence – Performace Evaluation

Figure 1.1 : Organization of Thesis

1.5 THESIS ORGANIZATION

Figure 1.1 illustrates the organization of this thesis.

1.5.1 Chapter 2: State-of-the-Art : Human Action Analysis

This chapter provides an overview of methods dealing with representation, classification and segmentation of human action videos. Such techniques are key ingredients for building an action assessment system.

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

This chapter covers the past work addressing human action assessment. Past work has been categorized based on its general applicability or task specific applicability for human action scoring and skill determination task.

1.5.3 Chapter 4: Assessment of Grace and Consistency

Long term actions like Sun Salutation or Warm Up Exercise is 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. In this chapter, we develop a framework to assess the pace and consistency of a performer. The framework provides feedback on where a wrong transition pace is taken by the performer. Our approach works by training individual Hidden Markov Models over spatio-temporal features for each dynamic posture. This enables automatic segmentation and labeling of the entire Sun

Salutation sequence using a concatenated-HMM. A modified Viterbi decoding algorithm is proposed in order to get a smoothened action label sequence. The timings for each dynamic posture as rendered by multiple experts are analysed to develop a metric for pace assessment which is then used 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.

1.5.4 Chapter 5: Detecting Missed and Anomalous Action Segments

While performing a long term action sequence, amateur performers tend to miss or wrongly perform a part of some action. Hidden Markov Models discussed in the previous chapter are strictly sequential in nature and fail to decode performances with missed or anomalous actions. Further, Dynamic Time Warping as a template-based matching technique also fails to align the sequences in presence of missed and anomalous segments. This is due to its boundary constraints. In this chapter, we propose an exemplar based Approximate String Matching(ASM) technique for detecting such anomalous and missing segments in action sequences. The technique involves comparing the performed sequence with the benchmark action sequence (as given by experts) and notifying when misalignments occur. For the evaluation of the proposed method, the Sun Salutation dataset is extended to include amateur performers who missed different intermediate poses in their performances.

1.5.5 Chapter 6: Assessment against multiple experts

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. Thus while the template based approach of finding anomalous sub-actions is a simplest approach to find discrepancies in an action sequence, different expert templates would give different feedbacks for the same test performance.

To overcome variations in feedback, we develop 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 sequences, the reconstructed sequences are similar to an expert rendering. Also, 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. Variations between the input video and the reconstructed video are exploited to provide appropriate feedback.

1.5.6 Chapter 7: Unsupervised anomalous pose detection

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 when the test sequences can have anomalous poses too. A technique that can naturally cluster the human poses without specifying the count of key poses is required.

We propose 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 results 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.

1.5.7 Chapter 8
Olympics Events Scoring relative to expert performance

Unlike Sun Salutation, the expert performances (which carry a high rating) are too few in case of Olympics events like diving and vaults. Thus the previously proposed autoencoder-based assessment technique is unsuitable for such a usecase. To overcome this limitation, we adopted a deep metric learning method that learns to score the similarity in performances of two input videos where the performance in the pair is not constrained to contain an expert performance. We propose a Long Short Term Memory (LSTM)-based Siamese network that learns discriminative features that are more relevant for predicting the score differences over pairs of training videos. The score prediction model learns to regress the concatenated embedding of Expert and a candidate performance to the ground-truth judge's score. This enables interpretability to given score, as one can capture and asses the temporal variation from given reference video. The proposed network is also used to predict the contributions of individual clips of the action over its final score.

1.5.8 Chapter 9
Conclusions and Future Scope

This chapter summarizes the research work done in this thesis and provides a conclusion with useful insights about human action assessment task. Some of the open challenges and possible future extensions are also discussed in this chapter.

...