4

A framework to assess Grace and Consistency of Long-Term Repeatable Actions

Physical exercises like Sun Salutation, Aerobics or Warm Up Exercise are long-term actions that constitute multiple poses performed in a sequence. These exercises are repetitive in nature and are required to be done with perfection to derive maximum benefits. The sub-actions or the inter-pose transitions should be performed smoothly with minimum jerks. Essentially, it is important that these exercises are performed with Grace and Consistency. By grace we mean that the approach and departure from a rest posture should not be very quick. A faster approach and departure from rest posture can be characterised as sudden movement or jerk, and it essentially deprives the benefit of the exercise. Consistency measures the repeatability of an exercise in every cycle. We propose an algorithm that assesses how well a person practices an exercise in terms of grace and consistency and test our algorithm for Sun Salutation dataset. Our approach works by training individual HMMs for each *asana* or sub-action using STIP features [Laptev [2005a]] followed by automatic segmentation and labeling of the entire Sun Salutation sequence using a concatenated-HMM. The metric of grace and consistency are then formulated in terms of posture transition times learned from expert performers. The assessments made by our system are compared with the assessments of the yoga trainer to derive the accuracy of the system. We introduce a dataset for Sun Salutation videos comprising 30 sequences of perfect Sun Salutation performed by seven experts and used this dataset to train our system.

4.1 INTRODUCTION

Assessing the quality of actions performed by the people has many important real-world applications. For eg. Yoga exercises have been included in the daily routine by many people now. Daily yoga practice is incomplete if not done in a proper way. A number of yoga researchers have emphasized on the rightful way the exercise should be executed to attain maximum benefits from it. The automatic assessment of yoga actions can provide the individuals a feedback on where they go wrong so that they can improve to attain benefits from their daily exercises even without the yoga trainer being present with them everytime.

Exercises like Sun Salutation, Warm Up Exercise and Aerobics are repetitive in nature and constitute a sequence of poses. Sun Salutation as an example is a sequence of ten subtly powerful postures set in a dynamic form performed in a single, conscious, graceful flow. This sequence is not performed once but is repeated many times where every cycle is consistently performed. For such exercises, the assessment based on poses of a person is incomplete and the between posture motion dynamics also need to be judged.

Omkar [2012] analyzed the grace and consistency of Sun Salutation using Inertial Measurement Unit sensors mounted on the human body and analysed non-linear signals obtained from them using signal processing techniques like Fast Fourier Transform and Wavelet transform. However, a yoga expert can assess the grace of a person by visually looking at how he or she practices the Sun Salutation. Thus a video analysis based technique has all the necessary information for the assessment of a Sun Salutation sequence based on the parameters of grace and consistency.

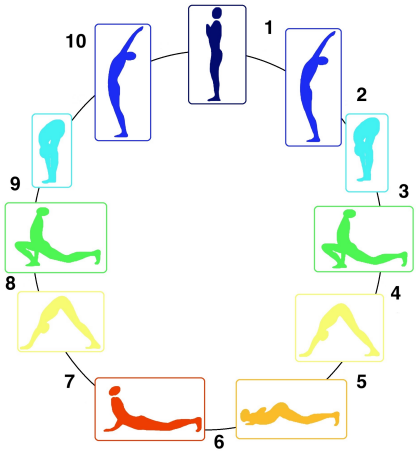

Figure 4.1 : Sun Salutation Sequence

Grace is defined as performing actions without jerks or sudden motions. In Sun Salutation, jerks can be considered as attaining a posture in a very short time duration and hence would appear jerky. Consistency on the other hand is the measure of repeatability of every action in each cycle of Sun Salutation and it can be judged based on how similar the time duration of the same posture were across different cycles.

We propose a framework that assesses how well people practice Sun Salutation in terms of grace and consistency. Our approach works by training individual HMMs over STIP features [Laptev [2005b]] for each dynamic posture followed by automatic segmentation and labeling of the entire Sun Salutation sequence using a concatenated-HMM. The metric of grace and consistency are then laid down in terms of posture transition times obtained from expert videos. The assessments made by our algorithm are compared with those given by the yoga trainer to derive the accuracy of the system.

A major technical contribution of this work is a framework for analyzing the sub action timings of the Sun Salutation sequence for assessing the grace and consistency. We propose a modification to the Viterbi decoding process in order to get a smoothened action state sequence. We release a new dataset for Sun Salutation assessment in hopes to facilitate research on this task.

For assessment of grace and consistency, our main requirement is an automatic recognition and segmentation algorithm that can define boundaries of the 10 dynamic postures of Sun Salutation and correctly identify them.

The baseline approach for action classification is the bag-of-words approach. In this approach, the multi-dimensional descriptors are first quantized based on a learned codebook. Then, for each action instance a histogram is computed over its quantized descriptors and used as input for a supervised classifier like Support vector machines. This approach has provided remarkable action-recognition accuracy [Laptev *et al.* [2008b]; Schuldt *et al.* [2004b]; Gilbert *et al.* [2009]]. Segmentation from Bag-of-words approach can be obtained by a windowed approach by splitting the video in overlapping windows and labeling each frame of the window by the action identified for the window [Duchenne *et al.* [2009b]]. The size of the window and the amount of overlap is completely arbitrary and choosing a correct value for the window size is difficult and is not capable to handle all temporal resolutions or time warped actions. Temporal graphical models is a more principle approach to segmentation models [Ozkan *et al.* [2012]; Borzeshi *et al.* [2013b]]. We adopt Hidden Markov Models to model the 10 sub-actions of Sun Salutation.

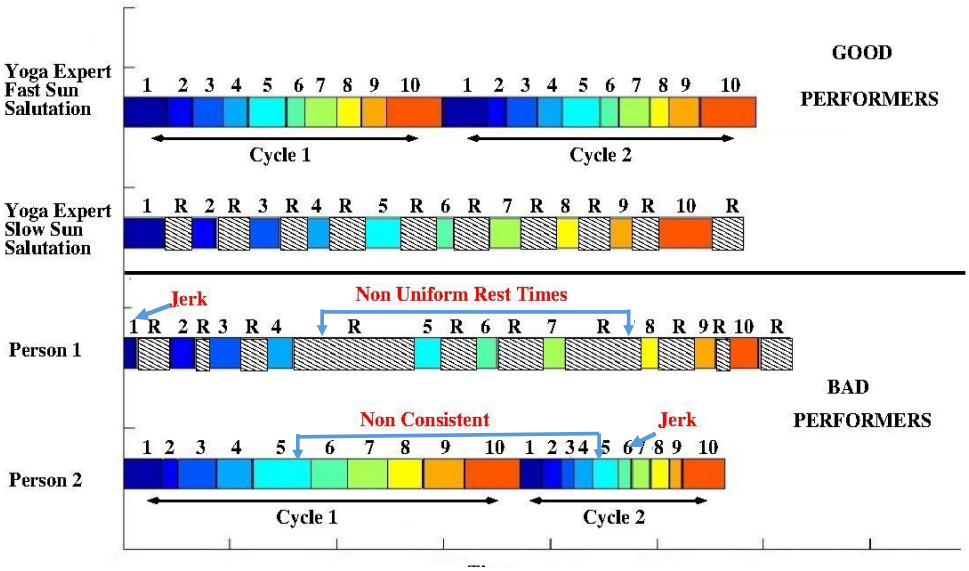

Figure 4.2 : Timeline illustrating jerks, rest times and consistency. From top to bottom sequence 1) 2 cycles of fast Sun Salutation by Yoga expert with no jerk, no rests, and consistent action. 2) 1 cycle of slow Sun Salutation by Yoga expert without jerk and uniform rest times 3) 1 cycle of slow Sun Salutation with jerk and non-uniform rest times 4) 2 cycles of inconsistent Sun Salutation with jerks

Since the primary requirement of this work is action segmentation, we model motion dynamics of a person using spatio-temporal features that have achieved good performance in action recognition task [Efros *et al.* [2003a]; Schuldt *et al.* [2004b]; Niebles *et al.* [2008b]]. They consider an activity in the 3D space-time volume, consisting of concatenation of 2D spaces in time. We use the shape/motion descriptors HOG-HOF [Dalal and Triggs [2005]] at spatio-temporal interest points [Laptev [2005b]], which are the locations of discontinuity in both space and time. These features are invariant to illumination changes, variations in view-point, occlusions, background clutter, and human clothing. Next we discuss the framework for assessment of motion dynamics of Sun Salutation sequence.

4.2 SUN SALUTATION : OVERVIEW

Sun Salutation or '*Suryanamaskar*' is one of the oldest yoga practices which has been practiced widely by people. The key postures are shown in Figure 4.1. We state some of the facts about Sun Salutation.

1. There are mainly 2 variations of Sun Salutation - fast Sun Salutation and slow Sun Salutation. The fast Sun Salutation is a yoga practice where the person transits from one key pose to another and immediately after reaching that key pose, he starts transiting to the following pose. Slow Sun Salutation requires a person to stay in a pose for some fixed counts after he attains that pose. We call the time taken to move from one pose to another as the "transition time" and the time a person stays at the same pose as the "rest time".

2. The transition times for different poses may vary according to the difficulty of a pose. Further, it is also person dependent. It is roughly between 1-2 sec. The execution must be rhythmic in nature, with each posture and its transition being executed with minimal jerks or ungainly movements. Thus the exercise needs to be performed gracefully.

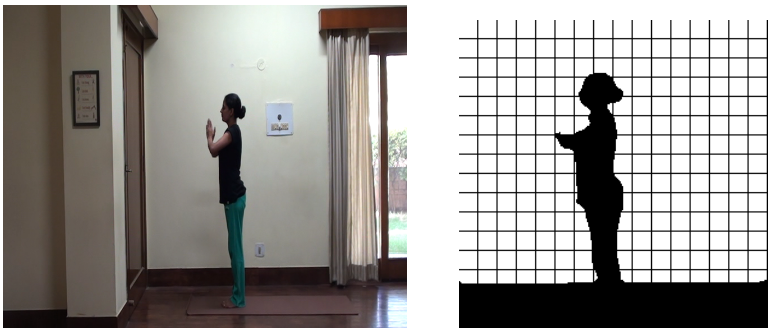

Figure 4.3 : Mesh features for a posture of Sun Salutation

3. The rest times at each pose in case of slow Sun Salutation need to be uniform throughout a cycle. However, the exact rest time is upto the person to choose. Some can choose it to 5 counts of breathing while others can take it to 7 and so on.

4. Sun Salutation is repetitive in nature i.e. is performed multiple times, and it is important that the same pace is followed. This leads us to another parameter of judging, which is Consistency - the ability to perform a number of cycles of the exercise, such that, each cycle is identical to another in the execution.

Essentially, we need to keep three main things in mind apart from gaining the correct postures - avoid jerks, have all rest times equal and be consistent between cycles. Thus, we analyze motion patterns of Sun Salutation sequences and check the correctness of the transition times of postures, rest duration and consistency between the cycles. Figure 4.2 illustrates the concept of grace, consistency and uniform wait times using timeline of two bad performers and yoga experts.

4.3 PROPOSED FRAMEWORK

We now discuss the framework for assessment of grace and consistency of Sun Salutation dataset. We first discuss the feature selection for this work, followed by individual sub-action modeling. This is then followed by segmentation and recognition of individual sub-actions in the entire sequence where we discussed our new Viterbi Decoding algorithm. Finally we discuss how to use the inter-pose transition times to identify jerks and inconsistency in the performance.

4.3.1 Choice of Features

We analyze two sets of features - *mesh features* which are shape-based features and *Space time interest points (STIP) features* which capture the pixel level motion of the body at every time frame.

Yamato *et al.* [1992b] have shown that the mesh features perform well for classification of various tennis strokes using Hidden Markov Model (HMM). Figure 4.3 shows the mesh features for one of the Sun Salutation poses. The mesh feature is extracted in three steps - i) the human detection algorithm extracts the bounding box of the human in every frame ii) the bounding box is resized to 256*256 and background extraction is applied to this resized image to give us a foreground mask where the black pixels represent the foreground and the white pixels represent the background iii) from every non-overlapping 16*16 cell in the resultant image, the ratio of foreground to background pixels is concatenated to form the mesh feature.

Mesh features vary according to the body configurations of a person. Thus to make an HMM

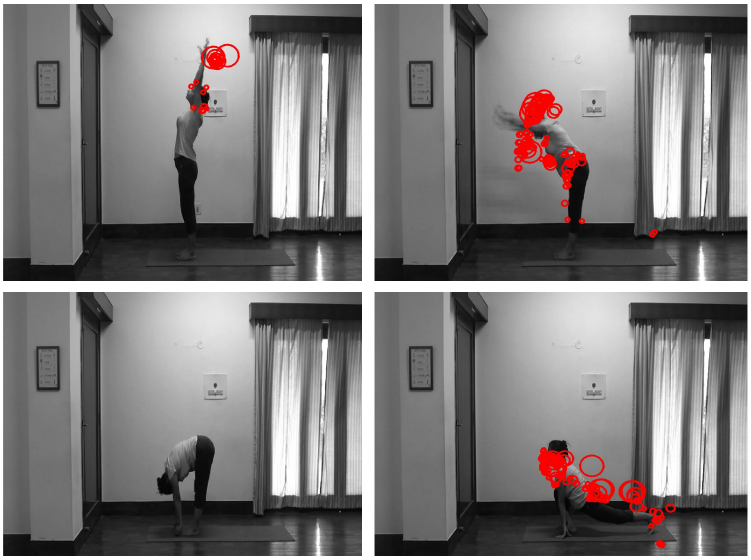

Figure 4.4 : Sample frames of Sun Salutation with variable no. of STIP features

adapt to multiple body configurations, we require the system to be trained with multiple people. Because we had very few yoga experts whose video could be used to train the system, the mesh features did not give proper results (as discussed in Experiments Section). Another drawback with this feature is that the foreground subtraction becomes difficult when the background is too cluttered.

STIP features [Laptev [2005b]] or 3D Harris detector detect regions having high intensity variations in both spatial and temporal dimensions (i.e. spatiotemporal corners). Interest points are detected for a fixed set of multiple spatio-temporal scales. We choose three scales to detect these features- 4, 8 and 16. For each interest point, we compute two patch descriptors - (i) histograms of oriented (spatial) gradient (HOG) - that describes local appearance and (ii) histograms of optical flow (HOF) that describes local motion. Thus for each detected feature point in all the frames we have a 162 dimensional feature vector (90-dimensional HOF and 72-dimensional HOG).

These features not only show better performance than mesh features for recognizing Sun Salutation actions with few videos of yoga experts, but they have two main added advantages which make us use it in our analysis. These are:

1. STIP features signify the locations of discontinuity in both space and time and therefore do not require any pre-processing such as background subtraction. Hence they can be used for any kind of videos irrespective of the background environment.

2. These features capture motion irregularities in the action. If there is no motion in certain time frames, no features would be detected and hence it makes it easier to identify the rest durations.

The recently introduced deep network architectures like Stacked Hourglass [Newell *et al.* [2016]] that provide human joints estimates are not fit for this application as the joint points in the consecutive frames may vary even if the person is in rest. Thus it predicts motion in a non-moving set of frames.

4.3.2 Sub-Action Modeling Using Posture specific HMM-MIO

Long-term actions like Sun Salutation is a sequence of 10 sub-actions where the motion dynamics are different as body transits from one posture to another, each sub-action/dynamic posture is modeled

using an HMM trained for that specific posture.

The STIP features that we use to capture motion dynamics of the actions exhibit irregularity from frame to frame. Figure 4.4 shows the result of applying the STIP detector and the corresponding variable number of STIP points. There are certain frames where there are no features detected while there are others where there are multiple features detected. Conventional HMM is not designed to deal with the variable number of observations per frame. To model such multiple irregular observations, the enhanced HMM, named hidden markov model with multiple, independent observations (HMM-MIO) [Concha *et al.* [2011b]] was developed as an extension to the traditional HMMs.

For multiple observations $O_t^{1:N_t}$ at each video frame t, Concha *et al.* [2011b] proposed the simplifying assumption of independence and identical distribution under a mixture. Here N_t denotes the number of observations at time t. They defined the observation likelihood for a state Q_t as:

$$p(O_t^{1:N_t}|Q_t) = \begin{cases} \sqrt[N_t]{\prod_{n=1}^{N_t} P(O_t^n|Q_t)}, & \text{if } N_t > 0 \\ 1, \text{if } N_t = 0 \end{cases}$$

Here the square root of the product of probabilities is taken to normalize the effect of the variable number of observations per frame. Equating $p(O_t^{1:N_t}|Q_t) = 1$ in case of no observations is equivalent to a missing observation and has neutral effect on chain evaluation of HMM. The forward and backward algorithm of traditional HMMs then proceed in a usual manner handling the multiple observations. We use HMM-MIO to model each individual posture. Two of the main architectures for HMM are - left to right and ergodic model. A left-to-right model involves fewer parameters and therefore is easier to train. Ergodic model on the other hand has increased degrees of freedom and tries to account for more observation sequences. In this work, we tested both types of models with various numbers of states to find the best structure to model the postures.

Another important concern is the choice of the number of states for every sub-action. We found that different postures may be best modeled by different number of hidden states, with more states for the more complex sub-actions. Thus to improve the model we also infer the optimal number of hidden states to be chosen for every posture.

4.3.3 Automatic Segmentation and Recognition of postures in long-term action sequence

Once we have the best HMM structure to model the individual segmented postures, we then need to segment the entire Sun Salutation sequence into different postures.

In continuous speech recognition [Martin and Jurafsky [2000]], the concatenation of HMM has been used for representing the phonemes in conjunction with the use of grammar. The phonemes in Sun Salutation are the key postures and the grammar is the order in which these postures should appear as in Figure 4.1.

We use concatenated HMM [Ozkan *et al.* [2012]] (Figure 4.5) to model the stitched Sun Salutation motion sequence, where the total number of hidden states in the concatenated HMM is equal to the sum of all the hidden states in the individual pose specific HMM. The priors, mean and variance of the Gaussian of the hidden states is obtained by simple vector and matrix concatenation. The transition matrix is such that the blockwise diagonal of the matrix is the one-label HMM transition matrix and the count of transitions between the class labels are computed to find the transition probability from one HMM to the other.

For the test sequence the most likely posture sequence is obtained as an output of the Viterbi

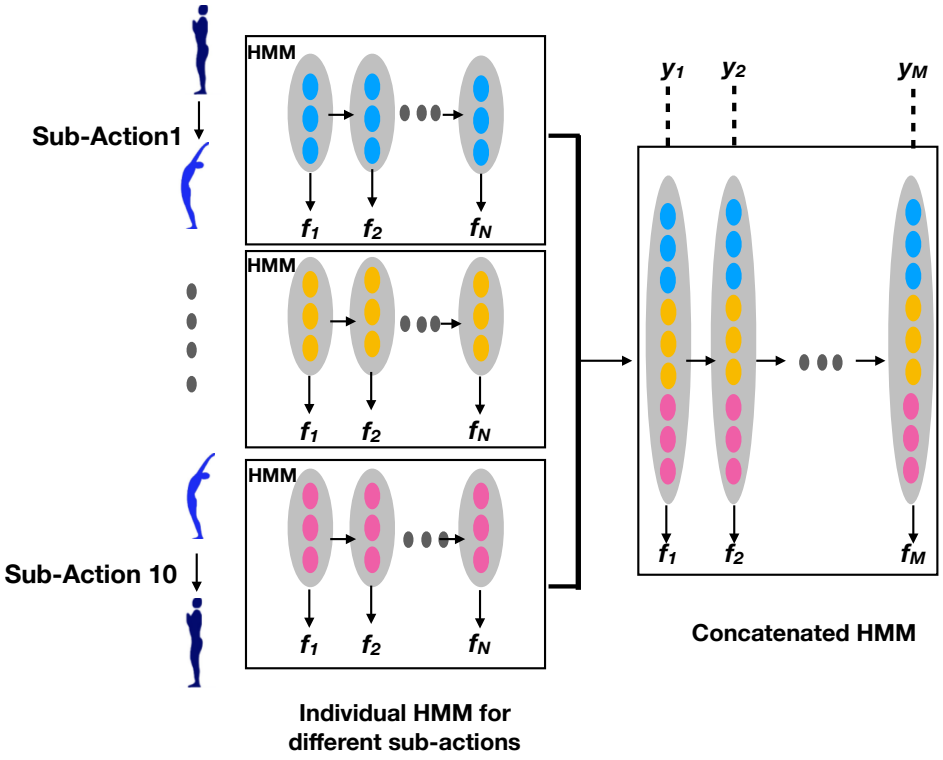

Figure 4.5 : Concatenated Hidden Markov Model for long-term action sequence

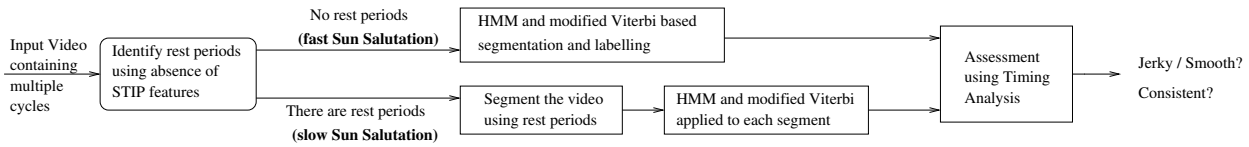

Figure 4.6 : Block Diagram for Fast/Slow Sun Salutation decoding

algorithm [Forney [1973]] executed over the concatenated HMM.

Our processing framework shown in Figure 4.6, applies different processing steps for fast and slow Sun Salutation sequences.

4.3.4 Decoding Fast Sun Salutation-Modified Viterbi with prior resets

We use the Viterbi algorithm for the concatenated HMM to produce most likely state sequence. The state that the HMM was in at time t corresponds to the action in the Sun Salutation sequence at time t. There are two main issues that occur in Viterbi decoded sequence.

Firstly, the Sun Salutation sub-actions are not distinctive in nature and at times are very similar in motion dynamics. The starting or ending portions of one dynamic posture may be similar to the start/end portion of another dynamic posture. For eg. the ending transitions of action 4 and 7 are quite similar to each other, as shown in Figure 4.7. This often confuses the Viterbi decoder and it outputs a quick changing state sequence. We show this with an example state sequence output of the Viterbi

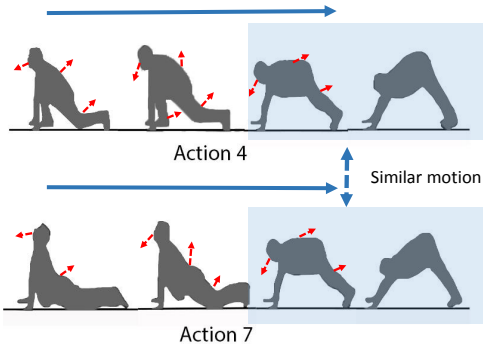

Figure 4.7 : Key poses of Action 4 and 7; Red arrows indicate motion direction of body parts

algorithm and compare it with the ground truth sequence.

Ground Truth Sequence

4 4 4 4 4 4 4 4 4 4 [4 4 4 4 4 5 5 5 5 5 5 5 5 5 5] 5 5 5 5 5

Viterbi Output

4 4 4 4 4 4 4 4 4 4 [5 6 7 7 7 7 7 8 9 10 10 1 2 3 4] 5 5 5 5 5

Here we see that because the ending transitions of both the actions 4 and 7 are similar, the Viterbi transits from 4 to reach to 7, which results in quick state changes involving the intervening states 5 and 6.

To make the classification result robust to undesired fast changes, Cohen *et al.* [2000] have introduced a smoothing pass of the state sequence in which they avoid changing the actual classification result if the HMM did not stay in a particular state for more than T times, where T can vary between 1 to 15 samples. Applying this smoothing with $T = 5$ we get the following output sequence :

Viterbi output smoothened by Cohen et al.

4 4 4 4 4 4 4 4 4 4 4 4 [7 7 7 7 7 7 7 7 7 7 7 7 7] 5 5 5 5 5

Here we see that since state 5 was occurring only once it got replaced with state 4 and so was state 6. However since 7 appeared for 5 times it remained the same in the smoothened output. Similarly, the states following 7 with frequency less than T were labeled as 7.

This output is free from undesired quick changing states. However the misclassified frames where action 4 is being labeled as 7 because of motion similarity are still present. To address this issue, we introduce a modification to the Viterbi decoding process to coerce it to produce smoothened state sequence. We reset the prior to the previous state if there is a change in state that does not stay for T time frames and then let the Viterbi algorithm re-initiate from that time frame in the usual way with the new prior.

We illustrate by considering the same state sequence by the unmodified Viterbi decoding

4 4 4 4 4 4 4 4 4 4 [5 6 7 7 7 7 7 8 9 10 10 1 2 3 4] 5 5 5 5 5

We observe that the state at index 11 i.e. action 5 has a run of length 1. Therefore, just before reaching state 5, we set the prior to 0 for all states other than those of action 4 and then let the decoding process proceed further. Viterbi decoding proceeds to give the following state sequence.

4 4 4 4 4 4 4 4 4 4 [5] 6 7 7 7 7 8 9 10 10 1 2 3 4 5 5 5 5 5

Here the underlined state denotes the index where the observation likelihood is more for action state 5 and inner boxed state is the one that is occurring for duration $< T$.

At index 12, the state 5 has run length 1, thus again we set prior to 0 for all states other than those of action 4 (the previous state).

4 4 4 4 4 4 4 4 4 4 4 4 [5 6 7 7 7 8 $\underline{9}$ 10 10 1 2 3 4] 5 5 5 5 5

The state sequence for the next 2 steps are :

4 4 4 4 4 4 4 4 4 4 4 4 4 [5 6 7 7 7 8 $\underline{9}$ 10 1 2 3 4] 5 5 5 5 5

4 4 4 4 4 4 4 4 4 4 4 4 4 [5 6 7 7 8 $\underline{9}$ 10 1 2 3 4] 5 5 5 5 5

After this step, as the Viterbi decoding is approaching the point from where confidence of state 5 would be high it transits to state 5 and we have the final sequence :

4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 5 5 5 5 5 5 5 5 5 5 5 5 5 5

The final output from the modified decoding process has got rid of the problems of frequent state changes and the misclassified chunks of frames. It is observed that the boundaries of actions 4 and 5 do not exactly match the input sequence. The state 5 begins at index 19 whereas in the ground truth sequence it begins at index 16. However, this minor lag does not much effect our subsequent analysis.

Ground Truth sequence

4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 5 5 5 5 5 5 5 5 5 5 5 5 5 5

Output

4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 5 5 5 5 5 5 5 5 5 5 5 5

Next we formally summarize the modified decoding process. Let A_j be the set of hidden states that belong to action j and X be the collection of subsets $A_1, \ldots, A_{10}$. For any $x \in X$, we define an indicator function :

$$\mathbb{1}_{A_j}(x) = \begin{cases} 1, & \text{if } x \in A_j \\ 0, & \text{otherwise} \end{cases}$$

The indicator function helps us to set priors for one action. Let $\mathbf{prior}_{All}$ be the vector formed by concatenating priors of individual action specific HMMs. For the concatenated HMM, the size of $\mathbf{prior}_{All}$ would be the sum of the number of hidden states in each HMM. We derive prior for an action j, while setting priors of all other actions to zero using :

$$\mathbf{prior}_{A_j} = \mathbf{prior}_{All} \odot \mathbb{1}_{A_j}$$

where $\odot$ denotes element-wise multiplication.

The modified decoding procedure as stated in Algorithm 1, makes use of the Viterbi decoded output.

The modified decoder keeps resetting the prior of the Viterbi output until the output is such that the decoded sequence has no posture making its appearance for less than *T* time frames. It is important to note that *T* is set to a small value of 10 so that even jerks are recognized. A larger value of *T* will force the posture to remain in valid ranges of posture timings and result in jerks not being recognized.

The Viterbi algorithm is prone to produce quick state changes when the observations can be attributed ambiguously to more than one action states. Resetting the prior has the effect of biasing

Algorithm 1 Modified Viterbi With Prior Resets

procedure VITERBI($Viterbi\_Decoded\_Output$)
  $[states, state\_run] \leftarrow$ RunLength$(Viterbi_decoded_output)$
  $Modified\_Viterbi\_output \leftarrow Viterbi\_decoded\_output$
  $index \leftarrow$ Find$(state_run <= T)$
  while $index >= 0$ **do**
    $state\_prev \leftarrow states(index(1) - 1)$
    Set $\mathbf{prior} = \mathbf{prior}_{A_{state\_prev}}$
    Re-initiate Viterbi with new prior from this point
    Update Modified_Viterbi_output
    $(states, state_run) \leftarrow$RunLength$(Modified_Viterbi_output)$
    $index \leftarrow$ Find$(state_run <= T)$
  end while
  return $Modified\_Viterbi\_output$
end procedure

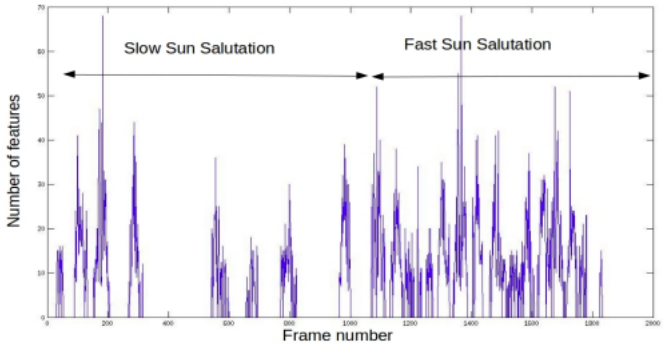

Figure 4.8 : Activity graph of a Sun Salutation sequence

the Viterbi output towards the hitherto more confident action state. This bias is repeatedly applied till the subsequent observation sequence can be attributed to a single action state which continues undisturbed for duration T or more.

Complexity of Modified Viterbi The complexity of traditional Viterbi decoding is $O(T \times |S|^2)$, where T is the number of frames in a video and $|S|$ is the size of state space S. The Modified Viterbi with prior resets has a complexity $O(k \times T^2 \times |S|^2)$, where $k \leq 1$.

$k = 1$ gives us the worst case complexity which means that at every time frame there is a confusion in action labels leading to prior re-initializations at every time instant. However, this is not the case and by experiments we found that once the prior has been reset, a Viterbi decoding iteration is able to set correct the labels of (some) more immediately following frames due to the involvement of transition probabilities. This kind of local progressive correction made by Viterbi has the effect of reducing the number of frames with label runs much less than T.

4.3.5 Decoding slow Sun Salutation sequences

Sun Salutation sequences generally contain multiple cycles of Sun Salutation where there can be non-uniform rest times between different postures. Figure 4.8 shows the activity timeline of a subject

who is not consistent in his rest times. Absence of features indicate rest time after a posture. The Viterbi algorithm described above fails to account for these variable size rest times and works well only in case of fast Sun Salutation. This is because for frames with no features detected, the observation likelihood of states is set to all ones by HMM-MIO, as discussed previously. For the consecutive sub-actions the end postures of the preceding action is same as the start of the following, and thus there are cases where before the rest times the likelihood of the observations gets high for the second instead of first. This leads to the incorrect labeling of the rest time frames entirely by the second action. To account for this problem, in slow Sun Salutation sequences we apply the Viterbi decoding process in segments. We consider the rest times as *neutral states* and for each activity segment preceded and followed by a rest segment, we decode the activity segment by the modified Viterbi algorithm described above. If the state at the previous frame just preceding the activity slot was j, the prior is set to $\mathbf{prior}_{A_j}$. This activity segment is then labeled by the action label which occurs in majority in the decoded output. The frames where the count of features is zero are labeled by the action label found in the previous activity slot. The above segmented decoding scheme is not applied to the fast Sun Salutation sequences because they don't have rest periods.

4.3.6 Grace and Consistency assessment

Grace - is defined as the performance of an action without sudden motions. This can be assessed by analyzing the transition times between the postures. The postures that are attained in too small time are the ones that are most accelerated and hence are considered as jerky. We need to determine a valid range for the transition time such that the actions which have transition times less than the lower bound are considered jerky.

The analysis of transition times in the videos of the Yoga expert suggests that the transition times of all the postures are different and depend on the difficulty level it takes to reach to the pose and is person specific too. However, there is a range in which the transition time lies which we have learned by examining the distributions of transition times in the sequences performed by the yoga experts.

We model the distribution of transition times of all the postures using a Gaussian. We consider the transition times in terms of the number of frames and it can be converted to seconds based on the frame rate which is 30fps.

To account for the possible mistakes in judgments caused due to the segmentation algorithm, we allow the variations of ± 30 frames, i.e. ± 1 second in the time while calculating the transition times. This is roughly equal to ± 1 standard deviation which constitutes 92.79% confidence interval. We find that the lower limit roughly comes out to be 32 frames or 1.06 seconds.

Rest time analysis - As stated in the general practicing norms of Sun Salutation, it is important that all the rest times are uniform in all cycles of Sun Salutation. For a test video we are unaware of how much time a person tends to rest. In addition to this there may be stretched or smaller rest times for certain postures. We take the median of all the rest time in a cycle and then compare all the posture rest times to this median again allowing 30 frames deviation. Any posture with a rest time more than or less than these values would be considered ungraceful.

Consistency - It is defined as the repeatability of the actions in every cycle. To assess this we compare the overall duration of a posture in the consecutive cycles, where

$$time = transition\ time + rest\ time$$

For the actions to be consistent these durations should be nearly equal within ± 30 frames. If the timings is postures in the consecutive cycles vary more than this we say the subject was inconsistent.

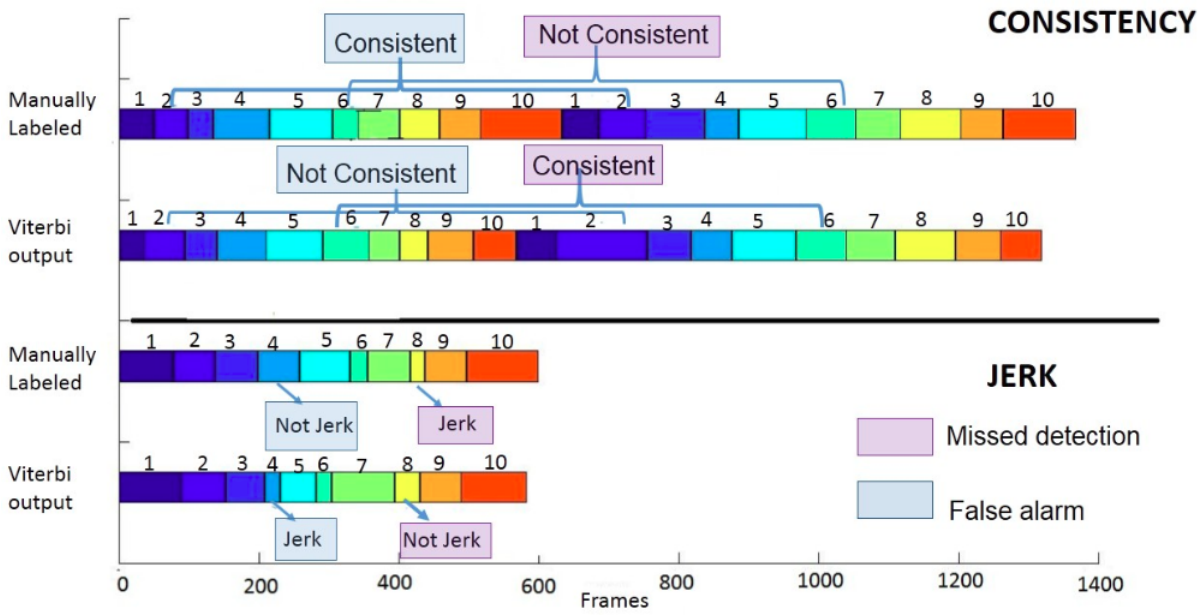

Figure 4.9 : Illustration of false alarms and missed anomaly detections

4.4 EXPERIMENTS

In this section we first identify the best structure of the HMM that we use for modelling the postures followed by the grace and consistency measurements. Since the Sun Salutation assessment has not yet been extensively studied in the computer vision community, we first introduce a new video dataset for Sun Salutation quality assessment available online at 'home.iitj.ac.in/∼jain.4/sun_salutation_assessment.html'.

4.4.1 Sun Salutation Quality Assessment Dataset

Training data - Sun Salutation videos were taken from 7 yoga experts who performed the Sun Salutation sequence multiple times. In all 30 Sun Salutation sequences were collected, which were then split into individual sub-action sequences for training the posture specific HMM. We observe that the transition times between dynamic postures are more or less consistent and lie in the range of 1 to 2 seconds

Validation data - Sun Salutation videos from 2 individuals were taken who performed Sun Salutation sequence four times. In all we had 8 Sun Salutation cycles, which were split into individual posture sequences. This was used as a test set to obtain the recognition accuracies of the trained HMMs of different configurations and find the optimal number of hidden states for each HMM to model the individual posture dynamics.

Test data - For assessment purpose, we recorded multiple cycles of Sun Salutation by 7 individuals. In all, we had 28 Sun Salutation cycles. The Yoga trainer judged these sequences on the parameters of grace and consistency. We used these judgements as ground truth for assessment. The videos were annotated by sub-action boundaries to evaluate the performance of our proposed modified Viterbi algorithm and the clips where the consistency and grace rules were not followed.

4.4.2 Results

Table 4.1 lists the performance accuracy of the mesh features and STIP features for 2 models of HMM - left to right with loopback to the first state, and ergodic HMM. The combination of STIP features

Table 4.1 : Performance accuracy of posture specific HMM

Model / Feature	Left-to-Right HMM	Ergodic HMM
Mesh features	37.5	45
STIP features	76.25	**80**

Table 4.2 : Comparison of segmentation accuracy of concatenated HMM with traditional Viterbi and the proposed Viterbi Algorithm

Viterbi Type / Model	Viterbi algorithm	Viterbi Algorithm with prior resets
STIP + Ergodic HMM	58.56	**81.516**

with ergodic HMM turns out to be the best combination to model the individual postures. We have utilized non homogeneous HMM, i.e. different number of hidden states to model different actions. Using the validation set we found that the optimal number of hidden states were two for actions 4 and 7 (see Figure 4.1), five for action 6 and seven for the remaining actions. Table 4.2 lists the performance of segmenting the Sun Salutation sequence using the Viterbi Decoder and its modified version.

Table 4.3 lists the accuracy of our assessment framework in terms of precision and recall for anomalies like jerks and inconsistent actions across different cycles. Figure 4.9 shows examples of false alarms and missed detections for jerks and inconsistent actions that reduce accuracy.

The performance of the Sun Salutation assessment on parameters of jerks and inconsistent actions depends on how well the segmentation algorithm segments the Sun Salutation sequence into sub-actions. In Table 4.9, we see that the precision-recall while analysing non-uniform rest times is 1. This is because the rest time boundaries are cleanly defined by frames where there are no STIP features detected and hence the rest time analysis is independent of segmentation errors. However, we observe that the primary contribution to the low precision (i.e. false anomalies) of 0.739 and recall, (i.e. missed anomalies) of 0.75 is the inaccuracy creeping into the action segmentation stage due to the HMM that leads to false stretching and shrinking of action durations.

Table 4.3 : Performance accuracy of assessment framework in detecting jerks and inconsistent segments

Anomalies	Precision	Recall
Jerks	0.75	0.75
Inconsistent actions across all cycles	0.739	0.894
Incorrect rest times	1	1

4.5 CONCLUSION

Sun Salutation is practiced rigorously by people in the present time. While practicing Sun Salutation it is important that it is done in a correct manner to derive maximum benefits. In this chapter, we propose an assessment system to judge the motion dynamics of Sun Salutation sequence performed by a person in terms of how gracefully a person performs the actions and how consistent he is across various cycles, assuming that the postures are all taken correctly. The sequence segmentation into individual postures is done using vision based techniques and the assessment is then made based on statistical comparison of the motion duration to those of a yoga expert.

Till now we considered mid-level performers who perform all poses in a sequence and may lag in their pace of action. Amateur performers on the other hand are prone to missing actions and perform anomalous segments. Hidden Markov Models learn pose transitions based on the training data. They lack in decoding sequences which differ in sequences of poses or have anomalous segments in between. Thus we need an algorithm that can handle assessments for amateur performers. In the next chapter we discuss template based approach to find missed and anomalous sub-segments in a performance

...