HUMAN ACTION ANALYSIS: NOVEL METHODS & PERSPECTIVES

A THESIS

submitted by

KARTIK GUPTA
(S14002)

for the award of the degree of

Master of Science
(By Research)

School of Computing and Electrical Engineering
Indian Institute of Technology Mandi
India - 175001

THESIS CERTIFICATE

This is to certify that the thesis titled **HUMAN ACTION ANALYSIS: NOVEL METHODS AND PERSPECTIVES**, submitted by **Kartik Gupta**, to the Indian Institute of Technology, Mandi, for the award of the degree of **Master of Science (By Research)**, is a bonafide record of the research work done by him under my supervision. The contents of this thesis, in full or in parts, have not been submitted to any other institute or university for the award of any degree or diploma.

Date:

Mandi, 175001

Dr. Arnav Bhavsar

Guide

Dedication

To my beloved parents, brother & sister

Declaration

I hereby declare that this submission is my own work and to the best of my knowledge and belief, it contains no material previously published or written by another person for award of any degree in any University, except where due acknowledgment has been made in the text. In addition, I certify that no part of this work will, in the future, be used in a submission in my name, for any other degree in any university. This thesis is a presentation of my original research work.

Candidates name and signature

Acknowledgments

First and foremost, praises and thanks to the God, for blessings throughout my research work to complete the research successfully.

I would like to express my deep and sincere gratitude to my advisor Dr. Arnav Bhavsar for his continuous support in research, also for his great help, patience and motivation. His guidance helped me in all the time of research and writing of this thesis. He was extremely patient with me and was always willing to discuss even the minute details.

Besides my advisor, I would like to thank Dr. Darius Burschka from TU Munich for his guidance and support in research and always ready to help behavior and immense knowledge I have taken from him while my research stay at TU Munich.

I am extremely thankful to Prof. Timothy A. Gonsalves, Director of IIT Mandi, for providing excellent research environment. I am thankful to Dr. Anil Kr. Sao, Chairperson of School of Computing & Electrical engineering for providing excellent facilities for conducting research work.

I am also thankful to the rest of my thesis committee: Dr. Renu M. Rameshan, Dr. Samar Agnihotri and Dr. C. S. Yadav for their encouragement and insightful comments.

Last but not the least, I would like to thank my parents for their immense belief on me, support and patience.

Kartik Gupta

Abstract

Automated human action analysis has important applications in various domains such as automated driving systems, video retrieval, video surveillance (for security purposes), elderly care, and human-robot interactions. However, various problems in this area are quite challenging and are yet unsolved. Traditional problem of human action recognition involves the classification of videos to action class labels. This requires a robust video representation technique and good classifier for modeling of feature representations and to account for variations. In real time applications, one has to deal with continuous action videos where multiple actions are performed. In some cases (e.g. human object interactions), one also needs to consider local levels of actions involving aspects of individual body parts and objects. In this thesis, we propose some approaches and provide some interesting experimental analysis to address some important problems related to human action analysis.

First, we propose to use skeleton information with Eigen-joint frame representation and apply a dynamic frame warping (DFW) framework and a Bag-of-words (BOW) framework for action recognition. Our approach can deal with the variations in action duration. We demonstrate that our method is better able to deal with the intra-class variations and as a result, performs better than some contemporary methods. Our approach also work with lesser number of training examples better than hidden markov models (HMMs) and conditional random fields (CRFs).

In the second part of the thesis, we consider a more challenging aspect of human action localization which is important for continuous action recognition. In this problem, a particular action is to be recognized in a test sequence of multiple actions, with unknown order. We do not assume any knowledge about the starting and ending frames of each action. We

propose a greedy alignment algorithm which works in real-time, and is extended upon the Dynamic frame warping framework. A notion of class templates in the DFW framework helps in achieving the intra-class variations and the greedy alignment algorithm allows us to work with framework in real time unlike dynamic programming based dynamic frame warping framework.

In the third part of the thesis, we focus on the task of fine-grained manipulation action classification where hand-object interactions are involved. In this work, we use grasp attributes and motion-constraints information available with Yale Human Grasping dataset. We propose to use the grasp and motion-constraints information to classify 455 object manipulation actions present in this dataset. We show differential comparisons for the performance of different classifiers on grasp information. We also compare object manipulation action recognition accuracies using coarse-grained and fine-grained grasp information.

Keywords: *Human action recognition, Human action localization, Object manipulation actions, Depth cameras, DFW, Greedy Alignment Algorithm, Grasp attributes and Motion-Constraints, RGBD.*

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Eon of evolution has gifted humans with advanced cognitive abilities to perceive and understand complex human activities and related behavior. This, in turn, enables humans to effortlessly interact with their surrounding environment. Such abilities involve understanding sequences of individual actions, and achieving significant invariance to changes in pose, illumination, etc. More specifically, the general task of activity analysis includes many sub-tasks such as temporal and spatial localization, action recognition, understanding human-object interactions, inferencing an intention of the action, predicting actions, etc.

The research in automated human action analysis considers problems of equipping machines to perform with some of the above capabilities. The challenges which make it difficult for machines to perform similarly like humans, are background clutter, partial occlusion, changes in scale, viewpoint, lighting, appearance, environment and understanding pose. Dealing with such challenges is interesting and hence, human action recognition problem has been studied since long and still a topic of great interest [6], [7]. Even after years of research in past decades, the task of human action recognition from the videos has addressed very simple scenarios such as restricted to isolated action recognition (where single video contains a single human action) and smaller set human action recognition (considering few action classes). Only in recent years, problems involving complex motion dynamics, human-object interactions, etc. have been considered in [8], [9], [10]. Another important motivation to consider such problems is their usefulness to applications in various domains such as au-

tomated gesture based control, video retrieval, video surveillance (for security purposes), elderly care, and human-robot interactions.

With progress in machine intelligence and automation, it now possible to control machines upto some extent with human gestures [11], [9]. It is also an essential part in the quest to make computers understand human actions and gestures in similar way as we perceive them. Eventually such a progression of machine from understanding simple actions to complex actions, to interaction with the environment, is important in leading machines to understand behavior in general. This thesis plays a small part in contributing to such a progression by addressing some contemporary problems such as human action recognition and localization.

Based on the complexity of actions, human action analysis can be categorized as:

- Human gesture recognition :- This problem basically consists of recognizing simple human gestures. Human gestures are elementary movements of individual body part (typically hands, face, etc.) and are the basic components describing the meaningful motion of a person. Some examples include sign language gestures or small movement of the person such as stretching an arm and raising a leg.

- Human action recognition :- This is more complex problem where a particular action is performed such as kicking, waving, walking etc. An action may consists of several human gestures combined to form an action.

- Human activity recognition :- Now the problem of activity recognition consists of understanding a complex activity that is performed by a human. An activity is composed of multiple actions combined to form a particular activity. For eg. playing football can be considered as an activity where several actions such as kicking, tackling are involved.

We consider problem which can be related to each of the above categories. First, we consider the human action recognition problem for the case of segmented videos where in each video there is single action class to be recognized. Although, there is lot of existing research to solve this problem as discussed in [6], [7], still the problem has quite challenging issues such as intra-class variations, temporal scale changes, etc.

Second, we address a more general problem where the video consists of multiple actions without information about the temporal location and duration of the actions. Also, the video stream is to be processed in real-time to consider the solution for real-time applications. The problem can also be considered as online action recognition.

For both the above tasks, we use skeleton data stream generated through Kinect and consider smaller set (5-20 classes) of actions. This is motivated by the usefulness of low-cost range cameras (e.g. Kinect) in the area of action recognition. Shotton et al. [12] proposed pose estimation algorithm which allows prediction for 3D positions of skeleton joints from a single depth image in real-time. Skeleton data stream generated through Kinect is typically more accurate and informative in terms of pose estimation and thus, is quite useful for the methods based on motion dynamics such as human action understanding. Also, there are inherent limitations of RGB data source, e.g. they are sensitive to illumination changes, color variations and background clutter. Range sensors give us 3D structural information of the scene and it's robust to the change in color and illumination.

Finally, we focus on a somewhat different problem than those considered above, which involves a larger set (455 classes) of hand-based manipulation actions. This can be related to the category wherein individual body parts are involved. However, unlike most works, we focus on actions involving interactions with objects. Everyday human actions involve hand-object interactions and there are subtle movement of skeleton joints in these actions. Thus, it is crucial to understand spatially local information associated with human hands rather than whole body pose to model these actions. Therefore, we use information about human grasp types and hand-object relations to recognize the actions associated with the objects.

1.1 Human action recognition

The traditional human action recognition problem is essentially a video classification problem. Here, each temporally segmented video contains a single human action and the problem is to classify each video among different action class labels. The assumption does away with

a tedious challenge of finding the temporal location of the action in the video. Nevertheless there can be challenges in this problem such as illumination and viewpoint variations if the videos are captured with visible-light cameras. Fig. 1.1 illustrates example frames of different actions in different scenarios. It clearly depicts intra-class variations and variation in illumination conditions for the same action class which makes even this simple problem difficult to deal with.

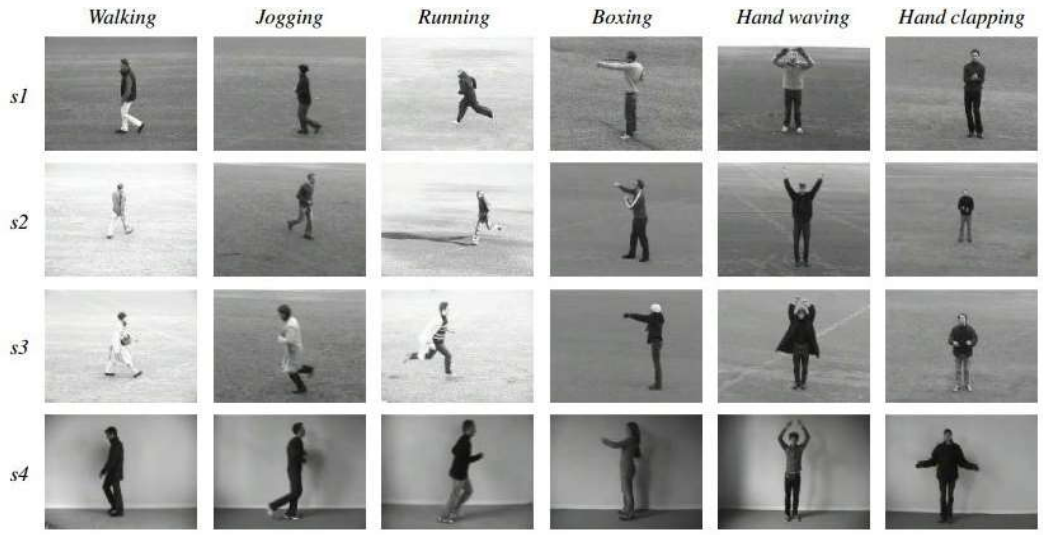

Figure 1.1: Examples of different actions in different scenarios (Reproduced from [1]).

As mentioned above, some of the challenges in visible light cameras are mitigated due to use of range cameras, as discussed above. We use skeleton information generated through Kinect. Skeleton data provides more compact information than depth data using pose based solution to different problems such as human action understanding.

Most of the methods targeting the problem of action recognition focus on video level features where each video descriptor describes a video with a single feature vector. This works well for the isolated action recognition i.e. action recognition for temporally segmented videos but fails to address a more real and relevant aspect of continuous action recognition. We have worked on frame-level features for isolated action recognition with a foresight of adapting the approach to that of temporal action localization in real-time continuous action videos. Frame level feature representation provides better information to model such complex

actions, where many skeleton joints are simultaneously moved. Apart from that, an action can be performed for different durations and there can be multiple ways, the same action can be performed with slight variations. The proposed framework uses a variant to dynamic time warping i.e. dynamic frame warping [13] which can align two temporal sequences of different lengths and incorporate intra-class variations.

1.2 Human action localization

Human action recognition in real-time applications requires to deal with continuous videos where multiple actions are performed by the human in an unknown order for unknown durations. The assumption of single action in a video does not hold in the real environment when dealing with the task of human action recognition. Unlike action recognition and offline action localization, which determine the action after it is fully observed, online action localization aims to localize the action on the fly, as early as possible. Thus, it is extremely important to consider the problem of action localization where we have a continuous stream of video for processing, being received by a system in real-time. Although online action localization is a significant problem, there are few works especially encompassing this problem such as [14], [15]. There are two important issues to be addressed in this task. First, we need to find the exact temporal location of any action. Second, we need to classify that action among the trained action class labels.

An illustration of action localization is depicted in the Fig. 1.2. It involves locating a particular action in the temporally unsegmented sequences consisting of multiple actions instances. We believe that this is one reason which makes this task more challenging than action recognition on temporally segmented sequences. Also, the problem of action localization involves considering the intra-class variations among different classes, viewpoint variations, and variations in temporal scale of the action sequence and the latency in localization of the action class.

Detection Latency can either be observational latency or computational latency as mentioned in [16], where observational latency is the time required by the system to observe

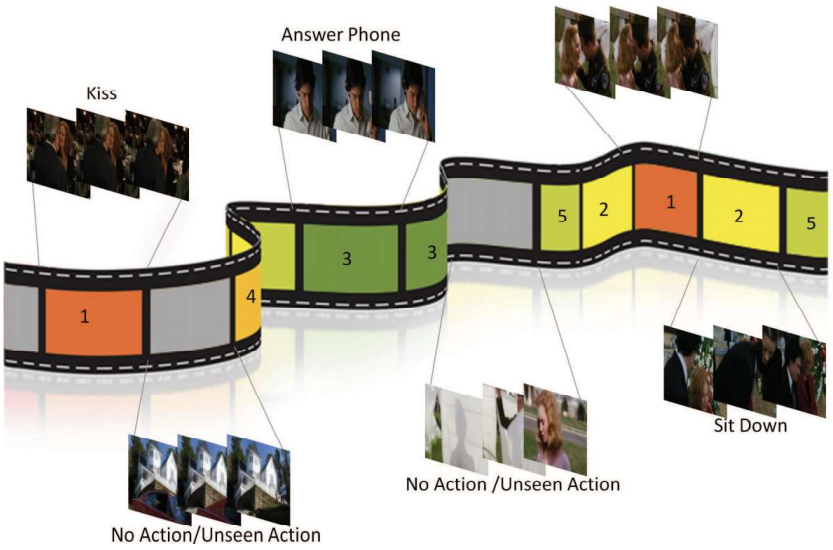

Figure 1.2: Realistic video contains multiple actions where the temporal locations and number of actions are unknown. Reproduced from [2].

enough frames to make a decision, whereas computational latency is the time required to perform the actual computation on a frame. Previous works [16], [17] & [18] either use a fixed temporal scale approach for the training and testing of the action class or they use a multi-scale approach. Temporal scale of the action can drastically vary for the same action and a fixed scale search strategy to localize the action in a continuous sequence cannot not work well. The use of video-level features to represent a video sequence instead of frame-level features, does not allow flexibility to change in temporal scale of actions.

Thus, an important aspect in the action localization problem is to deal with temporal scale variance at low localization latency. Our framework for localization works on the greedy alignment approach which has the capability to look for actions which may differ in temporal duration as compared to the samples used for training. The proposed framework has low computational and observation latency and on a standard machine with un-parallelized implementation, it can process 100 frames per second. Our action localization framework works in real-time with continuous stream of video being fed to it. This allows its scope for both video retrieval and real-time applications in surveillance, gaming, etc.

1.3 Fine-Grained recognition of object manipulation actions

An important challenge in human action analysis is that of considering human-object interactions. Although action recognition for the actions specific to the objects is a problem which has been studied in [19], [20], the recognition and understanding of everyday human actions is still difficult due to some hard challenges. There are subtle movements of skeleton joints in most of the manipulation actions. Also, the intra-class variations in manipulation actions are also significant. Thus, an action recognition approach which works on a large set consisting of hundreds of action classes is a very useful but challenging problem.

Some approaches [21], [22], [23] analyze human motion dynamics from video sequences. With the advent of cheap depth cameras like Kinect, the problem of action recognition has been dealt using motion trajectories [24]. However, Kinect body pose recognition readily fails when the user interacts with objects due to occlusion and limitations of depth based pose estimation.

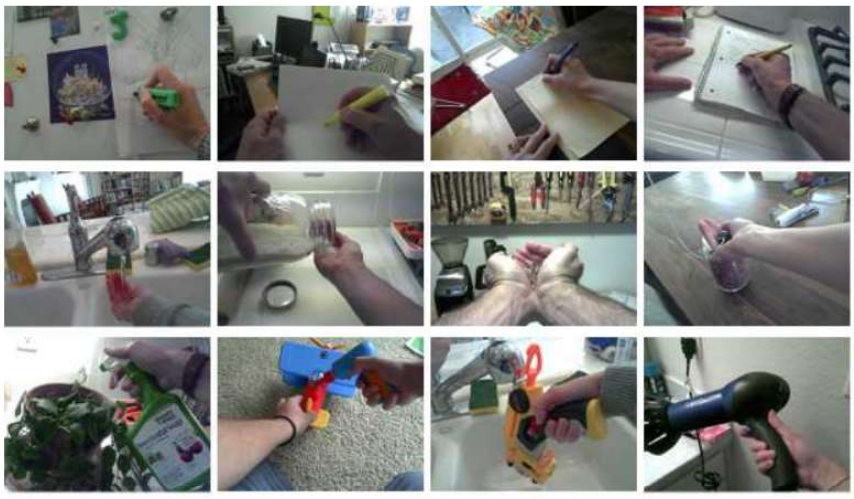

Figure 1.3: Sample frames of GUN-71: Grasp Understanding Dataset depicting variation in grasps for the objects.

Fig. 1.3 depicts some object manipulation actions from GUN-71: Grasp understanding dataset. The subtle manipulations which cannot be captured by the motion information

makes this problem difficult. This makes it important to find a solution other than human motion dynamics for the problem of action recognition. This motivates to look for solution to the task of human action recognition locally instead of globally. Considering spatially global information in object manipulation is not very useful as there are subtle movements for hands only. The spatially local understanding can use information associated with the hands of the actor in relation to the object in interaction. Humans have the ability to use their hands differently to accomplish the task intended with the object. Such information is particularly useful for the task of fine-grained recognition of object manipulation actions when there are hundreds of different actions being considered for task of classification. Our approach for fine-grained manipulation action recognition uses such spatially local information in the form of human grasp types at coarse and fine level.

1.4 Scope of the research

In this thesis, we investigate some contemporary problems related to the human action recognition domain.

We first address the problem of isolated action recognition using the Kinect generated skeleton data. The problem is still quite challenging due to issues such as intraclass variations, viewpoint variations, illumination conditions, temporal scale changes, etc. We use a dynamic frame warping framework which can incorporate intra-class variations unlike dynamic time warping. The class templates are modeled using various training sequences of each action through clustering of all training sequences with dynamic time warping (DTW) score. As mentioned above, our approach uses skeleton data information which is collected through Kinect.

Secondly for the action localization, we extend the earlier proposed DFW framework with a greedy alignment approach which has the capability to look for actions which may be smaller or longer in duration as compared to the training samples for those actions. The proposed framework has low computational complexity and on a standard machine with un-parallelized implementation, it can process 100 frames per second.

Both our works for action recognition and localization are proposed in aim to solve for the recognition of actions where motion information of the human skeletons is considered. We lastly use human grasp information and the hand-object relation to recognize the actions associated with the objects. The same object can be manipulated in different ways to perform a distinct action. Thus, we classify which action is performed on an object based on object information, human grasp information and other hand-object relational information. We propose this solution on set of hundreds of actions instead of small action set of 5-20 actions, typically considered in the above problems of human action recognition.

1.5 Contribution of this thesis

The contributions of thesis are as follows:

- We proposed to use dynamic frame warping framework (proposed for RGB in [13]) and bag-of-words framework to better model the intraclass variations using skeleton information from Kinect for the task of isolated human action recognition. DFW framework has an advantage of aligning two temporal sequences with different lengths and also can capture intraclass variations better using frame-level features.

- Extending upon earlier methods, we propose a novel greedy alignment algorithm using class templates as used in the dynamic frame warping framework to localize human actions in continuous videos from Kinect. This framework can work in real-time using skeleton information and has an adaptability to the temporal scale variations of actions.

- We propose a novel approach to fine-grained recognition of object manipulation actions. We use grasp attributes and motion-constraints to model the action information. We show recognition accuracies using coarse-grained and fine-grained grasp information. Our approach clearly highlights the usefulness of grasp attributes and motion-constraints for the task of fine-grained recognition of hundreds of object manipulation actions.

1.6 Organization of the thesis

The thesis is organized as follows:

In **Chapter 2** we propose a robust approach to solve the problem of human action recognition using depth cameras dealing with temporally segmented videos. This involves the dynamic frame warping (DFW) approach.

In **Chapter 3** we propose a novel real-time scale invariant human action localization approach. We show that the proposed framework using class templates from dynamic frame warping and novel greedy alignment algorithm, can deal with intra-class variations and temporal scale variations of actions. The greedy alignment algorithm proposed allows to run the action localization framework in real-time.

In **Chapter 4** we show the effectiveness of grasp attributes and motion-constraints information for the task of recognition of manipulation actions at coarse and fine level. Our results on Yale human grasping dataset consisting of 455 manipulation actions justifies our hypothesis.

In **Chapter 5** we summarize the contributions of the present work. We also present some directions for further work.

Chapter 2

Human Action Recognition using Depth Cameras

The problem of human action recognition is a challenging but important one, with applications in various domains such as automated driving systems, video retrieval, video surveillance (for security purposes), elderly care, and human-robot interactions. Traditionally, research in action recognition is based on video sequences from RGB cameras [7] or motion capture data [25], [26], [27]. Despite many research efforts and many encouraging advances, achieving good accuracies in recognition of the human actions, is still quite challenging.

The advent of Microsoft Kinect and similar depth cameras, have literally added a new dimension to the action recognition problem. Such low-cost depth sensors can provide both depth stream and the skeleton stream. In general, the problem of action recognition using depth video sequences involves two significant aspects to consider. The first is about effective representation of RGBD data, so as to extract useful information from RGBD videos of complex actions. The second aspect concerns developing approaches to model and recognize the actions represented by the suitable feature representation.

For the video representation, data from different modalities can be used such as RGB, depth or skeleton stream. As the skeleton stream generated through Kinect is very informative about the pose, we use an existing approach of skeleton joints representation, that of Eigen Joints [28]. The advantage of using this is that most of the existing works are mainly

on video level features but with Eigen Joints feature representation, we are able to work with frame level features which provides us more information and flexibility to work with.

Unlike the traditional RGB camera based approaches, the classification algorithm for the depth stream should be robust enough to work with small amount of training data, and handle intra-class variations as skeleton information is often noisy. In this respect, we explore a recently proposed work on Dynamic frame warping (DFW) framework for RGB based action recognition [13], for the task of depth based action recognition. This framework is an extension to Dynamic time warping framework to handle the large amount of intra-class variations which cannot be captured by traditional Dynamic time warping algorithm.

Unlike in [13], we do not use RGB features, but the skeleton joint features mentioned above. Some complex actions in skeleton based representations typically involve multiple joints movement simultaneously, which makes the problem harder. With more subjects performing same action in different environments in different ways, it becomes evidently important to come up with a more robust technique to deal with high intra-class variations. We consider different subsets of data, which highlight the above mentioned cases of complex actions and similar actions, and demonstrate superior performance of the proposed approach over the state-of-the-art.

To the best of our knowledge, such a dynamic frame warping framework on depth data has not been attempted till now. Such an adaptation from the technique proposed in [13], for action recognition in depth videos brings with it its own challenges. The formation of class templates in dynamic frame warping requires clustering of training sequences of an action class using dynamic time warping (DTW) distance. Motivated by this interpretation, we propose another approach, where we represent each video using bag-of-words (BOW) model, which is a clustering based model. Here, out of all the frames of training sequences, the most occurring frames are selected as cluster centroids.

We also note that, in conjunction with frame-level features, such a framework has another advantage over discriminative models like Support Vector Machines (SVM). It can be further extended as a dynamic programming framework, which can work for continuous action recognition in addition to isolated action recognition. Continuous action recognition involves

unknown number of actions being performed with unknown transition boundaries in a single video sequence. While, in this chapter, we do not consider the problem of continuous action recognition, such a problem is considered in the next chapter.

The experimental results clearly demonstrate that the proposed approach outperforms some of the existing methods for the cross-subject tests done on MSR-Action3D dataset [3] consisting of both complex actions, and actions with similar motion.

2.1 Related Work

With the advent of real-time depth cameras, and availability of depth video datasets, there is now considerable work on the problem of human action recognition from RGBD images or from 3D positions (such as skeleton joints) on the human body.

Li et al. [3] proposed a bag-of-words model using 3D points for the purpose of human action recognition using RGBD data. The authors used a set of 3D points from the human body to represent the posture information of human in each frame. In this work, the evaluation of approach on the benchmark MSR-Action3D dataset [3] shows that it outperforms some state-of-the-art methods. However, because the approach involves a large amount of 3D data, it is computationally intensive.

Xia et al. [29] proposed a novel Histogram of 3D Joint Locations (HOJ3D) representation. The authors use spherical coordinate system to represent each skeleton and thus also achieve view-invariance, and employ Hidden Markov models (HMMs) for classification.

In the work, reported in [28], the authors proposed an Eigen Joints features representation, which involves pairwise differences of skeleton joints. The skeleton representation proposed in this work, consists of static posture of the skeleton, motion property of the skeleton, and offset features with respect to neutral pose in each frame. The work involves Naive Bayes classifier to compute video to class distance. An important advantage with this representation is that it involves frame level features which not only captures temporal information better but also has an adaptability to continuous action recognition framework. Moreover, these features are also simple and efficient in their computation.

The approaches reported in [8], [30] & [31] also have been shown to perform well on the MSR-Action3D dataset. However, these works employ video level features instead of frame level features as we use in our work. We reiterate that with frame level features, it is relatively more straightforward to extend an approach for continuous action recognition, which is difficult with video level features.

In [13], Dynamic frame warping (DFW) framework was proposed to address the problem of continuous action recognition using RGB videos. Like the traditional DTW, this framework has the ability to align varying length temporal sequences. Moreover, an important advantage of this approach over DTW is that it can better capture intra-class variations.

Our proposed approach also uses the Eigen Joints feature representation, but in a modified Dynamic time warping framework as proposed in [13]. The major advantage with such a dynamic programming framework is it can work with frame level features, so it can understand the temporal sequence of frames better than Naive Bayes nearest neighbour classifier such as in [28]. In addition, as our experiments indicate, our approach can work without a large amount of training data required as in case of HMM (such as in [29]), as also indicated in [13].

2.2 3D Video Representation

As mentioned earlier, we employ the Eigen Joints features [28] which are based on the differences of skeleton joints. The overall Eigen Joints feature characterizes three types of information in the frames of an action sequence, including static posture, motion property, and overall dynamics.

The three dimensional coordinates of 20 joints can be generated using human skeletal estimation algorithm proposed in [12], for all frames: $X = \{x_1, x_2, ..., x_{20}\}$, $X \in \Re^{3\times 20}$. Based on the skeletal joints information, three types of pair-wise features are computed.

Differences between skeleton joints for the current frame: These features capture the posture of skeleton joints within a frame:

$$f_{cc} = \{x_i - x_j | i, j = 1, 2, ..., 20; i \neq j\} \tag{2.1}$$

Skeleton joint differences between the current frame-c and its previous frame-p: These features take into the account the motion from previous to the current frame:

$$f_{cp} = \{x_i^c - x_j^p | x_i^c \in X_c; x_j^p \in X_p\} \tag{2.2}$$

Skeleton joint differences between frame-c and frame-i (initial frame which contains neutral posture of the joints): These features capture the offset of an intermediate posture with respect to a neutral one:

$$f_{ci} = \{x_i^c - x_j^i | x_i^c \in X_c; x_j^i \in X_i\} \tag{2.3}$$

It is important to note here that, for some datasets it might be possible that neutral poses are not present in the initial frame of each sequence. For such a dataset, we can find a universal neutral pose for all the video sequences of training and testing sets of the whole dataset. We can find this one neutral pose from any video sequence of the dataset manually or by manually generating a skeleton with neutral pose and using its skeleton coordinates. To make this neutral pose comparable for the whole dataset, we need to normalize the skeleton locations, bone lengths and their viewpoints. To fulfill these requirement, we can translate the coordinate system such that hip center is at origin, rotate the skeletons such that vector from left hip to right hip is parallel to global x-axis and normalize body part lengths of all skeletons to make them equal to corresponding lengths in a reference skeleton. We do these normalizations because the skeletons may have variations in locations, viewing angle and bone lengths which may result in consistencies.

The concatenation of the above mentioned feature channels forms the final feature representation for each frame: $f_c = [f_{cc}, f_{cp}, f_{ci}]$. Feature rescaling is used to scale the feature in the range [-1,+1] to deal with the inconsistency in the coordinates. In each frame, 20 joints are used which result in huge feature dimension i.e. (190+400+400)*3=2970 as these differences are along three coordinates after feature rescaling, which gives us f_{norm}. Finally,

PCA is applied over the feature vectors reduce redundancy and noise from f_{norm} where we use leading 128 eigen vectors to reduce the dimensionality.

Such a feature representation on the depth videos is much more robust (in terms of invariances) than ordinary color based features on RGB counterparts of such videos, and also provide structural information in addition to spatio-temporal interest points.

2.3 Machine modeling of human actions via the Dynamic Frame Warping (DFW) framework

Dynamic frame warping is an extension to dynamic time warping and the overall advantage of using this framework is that it can deal with the intra-class variations. It also does not require a large corpus of training examples as needed in case of probabilistic graphical models.

2.3.1 Dynamic time warping (DTW)

Rabiner and Juang et al. [32], Mueller et al. [33] proposed Dynamic time warping (DTW) framework to align two temporal sequences $P_{1:T_P}$ and $Q_{1:T_Q}$ of unequal lengths. In this algorithm, the frame-to-frame assignments helps to match two temporal sequences:

$$A(P,Q) = \{(l_1, l'_1), (l_i, l'_i), \ldots, (l_{|A|}, l'_{|A|})\} \tag{2.4}$$

where $1 \leq l_i \leq T_P$ and $1 \leq l'_i \leq T_Q$ are indices of the frames of P and Q sequences, respectively.

The DTW algorithm finds the best alignment possible between the two temporal sequences P and Q. Each match between the elements of P and elements of Q gives a distance between that match while finding the best alignment. By passing through the best alignment path from $(1,1)$ to (T_P, T_Q), the matched distances are accumulated to come up with an overall DTW distance between the two temporal sequences as done in equation (2.5).

$$\text{DTW}(P,Q) = \frac{1}{|A|} \sum_{i=1}^{|A|} d(p_{l_i}, q_{l'_i}). \tag{2.5}$$

To find the best alignment path, a dynamic programming approach is used where the initial condition is $D(1,1) = d(1,1)$ and $D(p_{l_i}, q_{l'_i})$ is the cost for best alignment from start till the l_i frame of sequence P and $l_{i'}$ frame of sequence Q. The dynamic programming approach solves the following recursive solution illustrated in equation (2.6), by storing in memory the solutions to called functions.

$$D(i,j) = min \begin{Bmatrix} D(i-1,j) + d(i,j) \\ D(i-1,j-1) + d(i,j) \\ D(i,j-1) + d(i,j) \end{Bmatrix} \tag{2.6}$$

The final accumulated distance $D(T_P, T_Q)$ is normalized by dividing it by $|A|$, to find the final DTW score for two sequences P and Q. The normalization helps to overcome the variation in different number of matches for different alignments.

2.3.2 Dynamic frame warping (DFW)

As a variant to the traditional DTW algorithm, Dynamic frame warping i.e. DFW framework was introduced in [13]. This concept of DFW involves two main components: Action template represented by Y^l and Class template represented by $\tilde{Y}^l$ for each action class l. Here, the closest match to all the training samples of class l is found, $X^l_{i^*} \in \{X^l_n\}_{n=1}^{N_l}$. The closest match for each class l is defined as the action template of class l. Solving minimization in (2.7), yields the index of the sequence which is selected as the action template of each class:

$$i^* = \underset{i}{\operatorname{argmin}} \sum_{j \neq i} \text{DTW}\left(X^l_i, X^l_j\right) \tag{2.7}$$

Finally, denoting the action template of class l as Y^l, each training example X^l_j is aligned with Y^l using the above equation (2.7). This provides the class template:

$$\tilde{Y}^l = \left(\tilde{y}^l_1, \ldots \tilde{y}^l_{t'}, \ldots, \tilde{y}^l_{T_{Y^l}}\right), \tag{2.8}$$

with the length equal to length of Y^l, which constitutes of a sequence of metaframes. Each metaframe $\tilde{y}^l_{t'}$ is set of frames from the training sequences, which show a closest match with a corresponding frame of Y^l.