2

State-of-the-Art : Human Action Analysis

Human activity analysis has been an active field of research in machine learning community. It aims at recognizing activities of a person or a group of people using the acquired video data and the context where the action is being performed.

Human activity recognition (HAR) aims at interpreting events in the scene. From collecting the video data to analyzing what is being performed by the human in the scene, HAR systems involve the following steps – (1) *Feature extraction* - this involves extracting important representative spatial and temporal information from the videos (2) *Classification* - determining the the action class of a video (3) *Segmentation* – in case of long term actions, segmentation aims at localizing the constituting sub-actions.

A plethora of methods have been developed in the past few decades to address human activity recognition. The emerging popularity and importance of designing methods for human action recognition is attributed to provide users with automated assessment techniques that are useful in various application domains like health-care, sports and exercise monitoring, entertainment, activity of daily living, etc.

Traditionally, before the introduction of automated human activity analysis models, the task of analyzing the events in these domains was labor intensive and required around-the-clock human involvement. However, with the advent of machine learning approaches for human activity analysis, event understanding is being automated, thus benefiting the mankind.

Human activity assessment and monitoring involves comparing human activities with the acceptable behaviors and gestures performed for achieving specific goals in different domains. Here the preliminary step is activity recognition, followed by judging the performance. Examples of human action assessment systems include fall detection in health care monitoring application, posture-based human activity assessment in sports application, skill determination in kitchen, surgery, etc.

Different application domains demand different human motion representation and recognition techniques to meet the monitoring/assessment requirements. We report the techniques used in human motion representation and recognition, discussing applicability of these techniques in monitoring tasks and their shortcomings. We do not aim at reviewing the techniques in detail as it has been already done extensively in the previous reviews, rather we discuss the pros and cons of the techniques with the perspective of human action assessment.

In the next section, we discuss the previous reviews and how our review differs from them. Further, we describe some of the baseline HAR datasets to exhibit how the recognition techniques have progressed in the past and are able to handle complex human actions in the videos. In Section 2.2, we review the human motion representation techniques and their individual strengths and weaknesses. In section 2.3, we discuss various machine learning techniques used for atomic action encoding and recognition. Monitoring of complex actions and long-term activities requires monitoring at sub-action levels and thus need action segment boundaries to be determined. In Section 2.4, we briefly discuss the action segmentation techniques proposed in the past literature. Having laid the foundations of the task

of human action recognition, we discuss the past works in human action assessment in the following chapter.

2.1 REPORTED REVIEWS AND DATASETS

Human action recognition has significantly advanced in the last two decades and in the process, contributed to the advancement of level of activity datasets - from simpler to complex. In this section, we first describe the levels of abstraction for the human movement for the purpose of performing recognition. This is done to give the idea of basic terminologies to the readers and make the review easy to comprehend. Next, we list down the previous reviews towards human action recognition and discuss the evolution of publicly available action recognition benchmark datasets. We discuss how the progress in human action recognition systems has encouraged the researchers towards developing techniques for human action assessment and monitoring.

2.1.1 Human movement - levels of abstraction

The recognition of human movement can be performed at different levels of abstraction. We adopt a three level hierarchy of human movement - action primitive, action, and activity, as has been proposed by Moeslund *et al.* [2006]. Action primitive or gestures form the lowest level of hierarchy and are associated with movements at limb level. Actions can be simple and short-term or complex and long term and are described by the whole body movement that constitute action primitives. For example, walking is a simple action with repeated periodic action primitives while exercise sequences, dancing, aerobics are complex actions. Activities, on the other hand, consist of multiple actions performed in a defined order and in a set environment to achieve a task. "Moving one leg forward", "entering into the kitchen" and "preparing coffee" are examples of action primitive, action, and activity, respectively.

2.1.2 Reported reviews

The methodologies of human action recognition have been classified according to many different criteria in the past reviews. The criteria of types of action representation models (e.g. pose based features, volumetric, statistical) is used to classify the past works in literature [Forsyth *et al.* [2006]; Poppe [2010]].

Gavrila [1999] have used a taxonomy of 2-D approaches and 3-D approaches for action recognition. Moeslund *et al.* [2006] used 4 phases: initialization, tracking, pose estimation and recognition phases to discuss the past literature on human action recognition. They classified recognition techniques into holistic approaches, body part approaches and action primitive approaches. Later Turaga *et al.* [2008] addressed recognition problem under two categories - action, and activity, that vary in complexity. Weinland *et al.* [2011] surveyed action representation, segmentation and recognition techniques.

Aggarwal and Ryoo [2011] introduced four levels of abstraction for human action representation: gestures, actions, interactions and group activities. Human action recognition approaches were classified into two groups: Single layered approaches and Hierarchical approaches. The single layered approaches were used to recognize simple actions, while the complex activities were recognized using hierarchical approaches.

The reviews discussed so far have shed light on the contributions made by the computer vision research community towards developing human action recognition systems. The focus of these reviews are mainly on comprehending the human motion representation and the task of recognizing human movement at different levels (actions, activities, interactions, etc.).

The action analysis techniques have been developed to provide an intelligent, automatic,

interactive monitoring systems to the society. However not all representation and recognition techniques are useful in all domains, i.e. the human action recognition techniques developed so far are not generalized. If we choose any of the techniques, it may prove useful in some monitoring situations and fail to match the requirements of others.

While reviewing the past literature, we aim to address the strengths and weaknesses of the representation techniques and the recognition methods in order to meet the requirements of human action monitoring under different application domains.

2.1.3 Human action recognition datasets

There are many benchmark datasets proposed in the past for action recognition. These have been extensively reviewed according to their design, complexity and the intended purpose [Liu *et al.* [2011]; Chaquet *et al.* [2013]; Hassner [2013]]. We review some of these datasets to learn the evolution of human action analysis techniques from action level recognition to activity recognition that exhibits the growing capabilities of Computer Vision techniques towards human action analysis.

- **Early Action datasets** - **KTH** [Schuldt *et al.* [2004a]] and **Weizmann** [Blank *et al.* [2005]] are the early benchmark sets that have been extensively used over the years for comparing the performance of human action recognition techniques. These datasets include simple atomic actions like walking, jogging, boxing, bending, jump in place, etc. The videos in these datasets were collected in the lab setup with controlled conditions - static camera, uncluttered backgrounds, no occlusion of human and with well defined start and end action boundaries.

  IXMAS [Weinland *et al.* [2006]], a multi-view dataset was released to enhance for view invariant setting of human action recognition and with increase in the level of difficulty of action recognition task. This was again recorded under controlled conditions with minimal occlusions. All these datasets were limited in number of action categories (6 in KTH, 10 in Weizmann and 13 in IXMAS). The performance over these datasets have saturated over the years with action recognition accuracy over 95%.

- **Intermediate datasets** - Next generation of datasets were developed ignoring the controlled constraints like viewpoints and occlusions and were collected from TV and Sports sources. These videos are more realistic and challenging and hence increase the complexity of the recognition algorithms. A popular dataset under this category is the **UCF Sports** [Rodriguez *et al.* [2008]] dataset. It contains actions like diving, golf swinging, horse back riding, running, skating, etc. taken from various sports broadcasts.

  Feature films were also considered as a source of videos for human activity recognition. Two famous datasets in this category are **Hollywood-2(HOHA2)** [Marszalek *et al.* [2009]] and **Hollywood-1(HOHA)** [Laptev *et al.* [2008a]] benchmarks. It addresses verbs like answer phone, hug person, kiss, get out car, hand shake, etc. HOHA2 is an extension to the HOHA dataset. These datasets are special due to the fact that the actions in these datasets are not well localized in time. Thus the action recognition systems need to address the problem of *localizing* the actions in time too.

  The datasets discussed so far have included not more than 15 action categories. To make the task of classification even more complex a few datasets like **HMDB51** [[Kuehne *et al.*, 2011]] which includes actions like hand-waving, drinking, sword fighting, diving, running and kicking with cluttered background and large variations in camera viewpoint and motion and appearances of the actors and **UCF50** [Reddy and Shah [2013]] which includes actions like Baseball Pitch, Basketball Shooting, Bench Press, Biking, Biking, Billiards Shot,Breaststroke, Clean and Jerk, Diving, Drumming, Fencing, Golf Swing, Playing Guitar, etc., were released recently. This dataset

contains 50 action videos collected from various sources such as movies, YouTube and other public databases.

Finally **Action Similarity Labeling (ASLAN) dataset** [Kliper-Gross *et al.* [2012]] forms a benchmark human action dataset containing 432 action categories. Instead of designing a multi-class categorization algorithm, the ASLAN benchmark is used for designing algorithms for binary pair matching. The task is to decide if two videos present the same action or not. It encourages the development of action similarity measures and techniques, rather than developing methods to learn discriminative properties of actions.

- **Recent years - From simple actions to complex actions and activities** - Till now we have been discussing human datasets that are composed of atomic actions. In the recent years, an attempt has been made to move beyond actions, by considering activities involving several such actions performed together in a sequence, eg. preparing coffee, long jump etc.

  The **CMU Multi-Modal Activity Database (CMU-MMAC)** [De la Torre *et al.* [2009]] contains human activities acquired from multiple modalities of subjects performing the tasks involved in cooking recipes like brownies, pizza, sandwich, salad, and scrambled eggs. The videos taken were both egocentric and from static cameras.

  Olympics Sports dataset [Niebles *et al.* [2010]] was released to have 16 sports activities downloaded from YouTube. An example activity in this dataset is long jump that combines standing, running, jumping, landing and standing up again.

  The **Cornell Activity Dataset-120 (CAD-120)** [Koppula *et al.* [2013]] is composed of videos of 10 high-level activities like making cereal, taking medicine, stacking objects, microwaving food, picking objects, cleaning objects, etc. These activities are long sequences of sub-activities, that vary from subject to subject significantly in terms of length of sub-activities, order of sub-activities as well as in the way the task is executed. The sub-activity labels for the dataset are: reaching, moving, pouring, eating, drinking, opening, placing, etc

  UTK-CAP dataset [Zhang *et al.* [2014]] is a collection of videos acquired using five Kinect depth sensors. For example it has a sequence of actions recorded in a small gift store scenario, with action categories such as grab box, pack box, push box, use computer, write on board and answer phone. The long sequence of actions in this dataset also features gradual transitions between the adjacent activities, similar to the real life situations.

Human action recognition techniques have evolved over the years and so have the datasets which have progressed to include more complex action videos. In the past few years human action assessment has received the attention of researchers and has led to introduction of video datasets that not only describe the activity labels, but also the assessments such as human skill level, performance scores, etc. associated with these activities. We would discuss these datasets in the next chapter as we review the action assessment techniques.

2.2 HUMAN MOTION REPRESENTATION

The human motion can be represented by four different methods - by human body shape models, by human image models, by local feature methods, and by motion features from deep learning paradigm. We discuss all these methods in detail here and list their pros and cons from the perspective of human motion assessment.

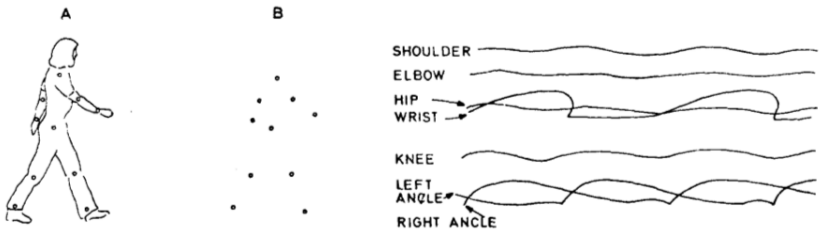

Figure 2.1 : a) Outline contour of a walking subject and their corresponding dot configuration b) motion paths of seven side joints of the walking person (reprinted from [Johansson [1973]] © Springer, 1973)

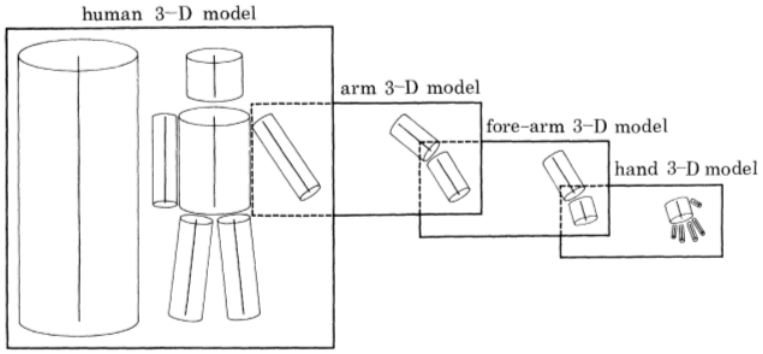

Figure 2.2 : 3-D model description of humans using cylindrical parts (reprinted from [Marr and Nishihara [1978]] © The Royal Society, 1978)

2.2.1 Human body shape models

This set of representation techniques models people using their joint positions in 2-D(X, Y) or 3-D(X, Y, Z) space. This kind of representation is motivated by two pioneering works from 1970's. Johansson [1973] recommended that by tracking the joint positions of the humans, we can distinguish between various actions that they perform. To validate this, they attached light displays(MLD) to the human body parts and validated that the trajectory of joint locations are sufficient to understand human motion (Figure 2.1).

Following this Marr and Nishihara [1978] conceptualized the perceptual recognition of objects and it was shown that the humans parse objects as a combination of general shape primitives like cylinders, cones, ellipsoids, etc. The sample human construction under this philosophy can be seen in Figure 2.2.

There are two set of techniques introduced in the literature, to record human joints. One set of techniques use the marker-based system to get the joint data while others use markerless techniques to estimate body joints and skeletons.

1. **Marker-based systems** - Early works in human joints tracking were based on magnetic sensors [Badler *et al.* [1993]; Molet *et al.* [1999]]. Though the output of the magnetic sensor-based systems is online and can be directly used to mimic human joints, these are discouraged for use for the reasons that they are prone to noise from the surroundings (eg. if a person is in close proximity

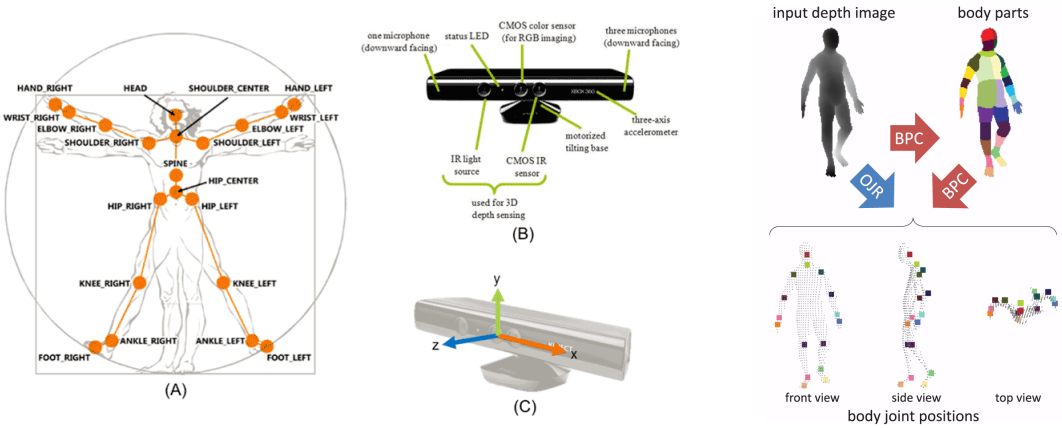

Figure 2.3 : a) Kinect camera with its camera configuration b) Pose estimation using a depth camera (reprinted from [Shotton *et al.* [2013]] © IEEE, 2013). Body part label at each pixel is predicted using Body part classification (BPC) and used to localize joints

with somebody else wearing magnetic sensors).

During SIGGRAPH in 1999, optical technology was demonstrated to provide better joint tracking capabilities. Motion capture systems with VICON cameras or Elite cameras [Ferrigno and Pedotti [1985]] were used to capture the motion data. In such a system, markers were attached to the human body and their locations were recorded by the cameras while the human performed actions. These marker positions are then fitted into skeletal joint trajectories. Herda *et al.* [2000] provides an automatic 3-D reconstruction process for robust skeletal tracking algorithms from the marker based trajectories.

Marker-based systems provide high accuracy in modeling human actions. However, their use is discouraged for the purpose of human action assessment due to the pain of attaching markers every time while performing the activity and the expenses of an 8-camera setup to read the positioning of the markers.

2. **Markerless techniques** - The high expenses of the marker based VICON system encouraged the researchers to develop less expensive markerless approaches towards tracking human motion. Markerless pose estimation techniques can further be divided into two categories - techniques using depth cameras, and techniques using kinematic models.

   - **Depth camera based approach** - With the advent of infrared sensors, the use of depth cameras is encouraged in the human motion data capture systems. These cameras are well suited to capture articulated human motion which is otherwise difficult to capture using monocular video sensors. The most popular depth sensor is the Microsoft Kinect camera. It produces depth, texture, user and skeleton information while capturing human motion data.

     Kinect cameras are equipped with two cameras - IR and RGB, that provide depth information and texture information respectively. The binary images provide the information about the detected people in the scene. A middleware framework *MS kinect SDK* provides us with the human skeleton in real time without calibration requirement. Two versions of Kinect camera have been released - Kinect v1 provides coordinates for 20 joint points of the human skeleton and Kinect v2 provides coordinates for 25 joint points. Thus Kinect v2 camera is more robust

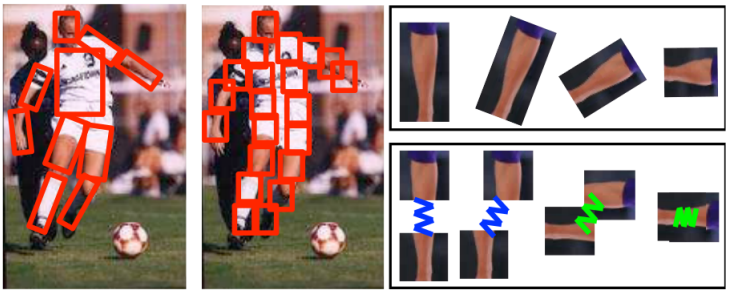

Figure 2.4 : Mixture-of-parts model [Yang and Ramanan [2011]; Felzenszwalb and Huttenlocher [2005]]; in the top-right model a single template is warped to different orientations. In the bottom-right, the body parts are Approximated using small warps by translating patches connected with a spring (reprinted from [Yang and Ramanan [2011]] © IEEE, 2011)

to the task of human tracking due to increased number of joint points (Figure 2.3). In addition to the Kinect SDK, there are several vision based techniques that use the depth sensors and extract human skeleton using the depth maps [Raptis *et al.* [2011]; Wang *et al.* [2012]; Shotton *et al.* [2013]].

Though Kinect cameras are cost effective and have shown good results towards joint tracking, they have some limitations too. Firstly, these cameras have a limited range (about 5 meters). The quality of pose estimation degrades as the distance of the person from the camera increases. This is due to noise and low resolution of the depth image. Secondly, skeleton tracking fails when occlusion occurs. Thirdly, the depth cameras are not readily available with people. RGB cameras as mobile cameras and surveillance cameras are instead more easily available with people. However, even with the limitations, it has been a first choice in clinical setup and is used in many rehabilitation scenarios [Chang *et al.* [2011, 2012]; Lange *et al.* [2011]].

- **Computer Vision based pose estimation technique** With the advent of computer vision research, configuration of body parts like limbs, head, trunk, etc. are modeled as kinematic tree. The tree constitutes joints with their corresponding links. All degrees of freedom(DOF) of joints in the body model, together, form the pose representation. The human pose estimation is divided into two phases - a modeling phase and an estimation phase. Modeling is the construction of the likelihood function for human pose representation, while estimation deals with finding the most likely pose. Poppe [2007] surveys techniques to model and estimate human pose from monocular images.

  Felzenszwalb and Huttenlocher [2005] introduced a framework that divides the appearance of an object into parts, with geometric constraints on pairs of parts. The pose estimation using these models use parts described by their locations only, thus simplifying the problem of inference and learning. However, poses are incorrectly estimated when a single template is warped into different limb rotations and foreshortening states. To address this problem, Yang and Ramanan [2011] used a mixture of small, non-oriented parts, and the flexible mixture model jointly captures spatial relations between part locations and co-occurrence relations between part mixtures (Figure 2.4). This leads to better pose estimation accuracy.

  Recently pose estimation models that use deep networks to estimate the poses in the frames have been introduced. DeepPose [Toshev and Szegedy [2014]] approach towards pose estimation is formulated as a CNN-based regression problem towards body joints where

the L2 Loss is minimized. They also use a cascade of such regressors to refine the pose estimates and get better estimates. Initial coarse pose is refined and a better estimate is achieved. Images are cropped around the predicted joint and fed to the next stage, such that the subsequent pose regressors see higher resolution images and thus learn features for finer scales which ultimately leads to higher precision.

Newell *et al.* [2016] introduced a stacked convolution neural network-based approach to pose estimation and has shown good results over the MPII [Andriluka *et al.* [2014]] and FLIC [Sapp and Taskar [2013]] datasets. Pishchulin *et al.* [2016] introduced DeepCut which is a bottom-up approach for multi-person human pose estimation. The authors approached the task of pose estimation by following steps: i) a set of 'D' body part candidates are produced. This set represents all possible locations of body parts for every person in the image and a subset of body parts from the above set of body part candidates are selected ii) each selected body part is selected from one of 'C' body part classes. The body part classes represent the types of parts, such as "arm", "leg", "torso" etc. iii) the body parts are then partitioned person-wise.

OpenPose [Cao *et al.* [2018]] is one of the most popular bottom-up approaches for multi-person human pose estimation. The OpenPose network first extracts features from an image using the first few layers of VGG-19. The features are then fed into two parallel branches of convolutional layers. The first branch predicts a set of 18 confidence maps, with each map representing a particular part of the human pose skeleton. The second branch predicts a set of 38 Part Affinity Fields (PAFs) which represents the degree of association between parts. Successive stages are used to refine the predictions made by each branch. Using the part confidence maps, bipartite graphs are formed between pairs of parts and using the PAF values, weaker links in the bipartite graphs are pruned.

The computer vision-based techniques to model human pose are advantageous because most of the videos, whether in daily life or in events like dance shows or Olympics or other sports leagues, are taken from RGB cameras and thus monitoring poses in such cases can be done with such techniques. However, these techniques still lack performance accuracy in case of occluded postures and completely unseen poses because the models learned are dataset specific and difficult to generalize. Further, low resolution video frames lead to incorrect pose estimation. A wrongly estimated pose can lead to wrong assessment results.

Pose features are also not suitable for human actions that involve objects such as a ball in the case of sports videos or in cases where the actions result in outcomes that do not contain human such as splashes in case of diving action.

2.2.2 Human image models

Human image models are also called holistic approaches i.e. humans are explicable only by the reference to the whole body structure, and the interconnections between the parts are not accounted. Silhouettes, contours are some representation techniques in this category.

Yamato *et al.* [1992a] were the first to use silhouette images to represent humans while performing action recognition. Human shaped mask was extracted for each image (Figure 2.5) and the ratio of foreground to background pixels for each cell in the grid combined to form the mesh feature.

Bobick and Davis [2001] introduced temporal templates for action recognition. They extracted human masks from video frames using background subtraction and accumulated the differences between consecutive frames. The difference in frames were used to construct a binary motion-energy image (MEI) and a scalar-valued motion-history image (MHI) where the former represented the

Figure 2.5 : Mesh feature (reprinted from [Yamato *et al.* [1992a]] © IEEE, 1992), sample shape masks for foreground stroke from tennis

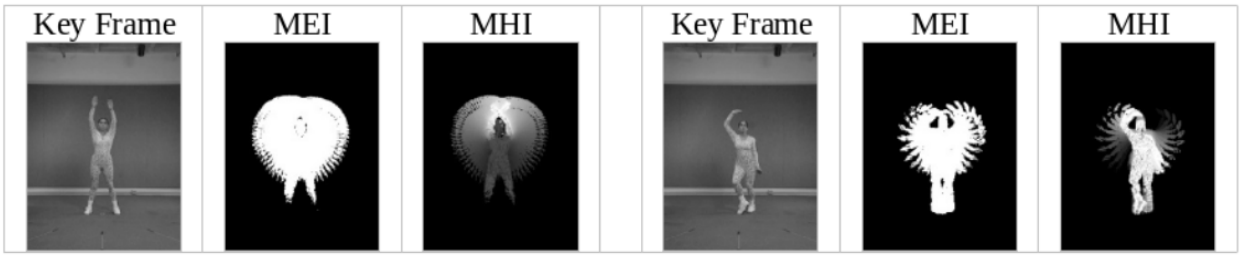

Figure 2.6 : Sample MEI and MHI images (reprinted from [Bobick and Davis [2001]] © IEEE, 2001)

presence of motion while the latter captures history of motion (Figure 2.6).

Blank *et al.* [2005] proposed three-dimensional shapes of human body silhouettes in the space-time volume (Figure 2.7). A spatio-temporal volume is formed by stacking silhouettes extracted at every frame using background subtraction technique.

Shape Context(SC) [Belongie *et al.* [2002]] of silhouettes is another representation technique used by Lv and Nevatia [2007] that gives a robust shape representation which is scale and translation invariant. The SC of each sampled coordinates from the remaining set of points is measured using the reference point as the origin. The representation is invariant to scale and translation as it uses relative scale and position. Lv and Nevatia [2007] used 12 angular and 5 radial bins were used for each SC and 200 edge points were sampled uniformly on each silhouette. (Figure 2.8).

Pros and cons of human image representation: The human image-based features are simple to compute. These features have shown good action recognition performance for actions performed in plain backgrounds. However for monitoring tasks, these features are of limited use as they do not accurately model humans as compared to body shape models such as pose features. Further, actions

Figure 2.7 : Stacked silhouettes from action frames (reprinted from [Blank *et al.* [2005]], © IEEE, 2007)

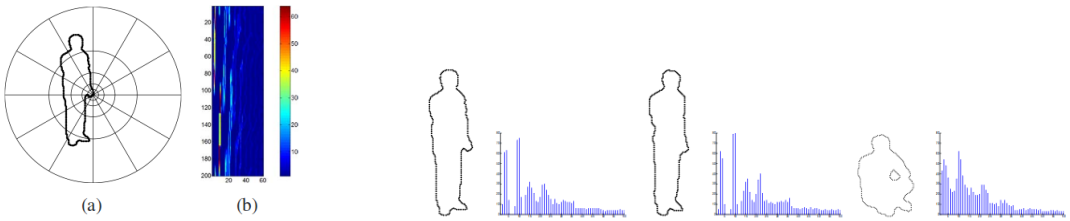

Figure 2.8 : a)Shape context feature for human representation (reprinted from [Belongie *et al.* [2002]], © IEEE, 2002), b)matrix representation of shape context. Reference silhouette with 2 query silhouettes

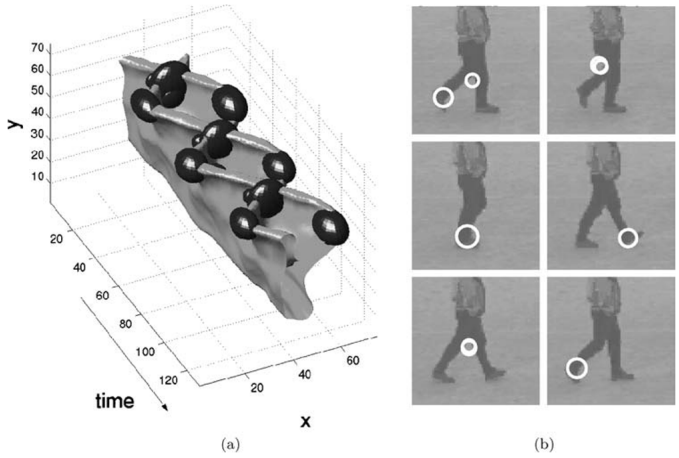

Figure 2.9 : Space time interest point features for a walking sequence (reprinted from [Laptev [2005a]], © Springer, 2005)

are mostly performed in cluttered background like kitchen, office, etc. and for such situations extracting human boundaries from backgrounds is noisy.

2.2.3 Local feature-based representation

The presence of background clutter, multiple motions and occlusions makes the tracking of part-based models difficult.

To overcome this issue, a number of local interest points(IPs) detectors have been introduced in the past. Space Time Interest Points(STIP) [Laptev [2005a]] (Figure 2.9) are obtained using the 3D Harris interest point detector, that computes a second-moment matrix at each spatio-temporal video point. The regions that have significant eigenvalues for the computed matrix are the ones that are positive local maxima. These define prominent motion points in the video.

Dollár *et al.* [2005] proposed the Gabor detector, that finds denser interest points as compared to Harris3D. The interest points are defined by the local maxima of a set of spatial Gaussian kernels and temporal Gabor filters response function.

Hessian3D detector [Willems *et al.* [2008]] is a spatio-temporal extension of the Hessian saliency measure used for blob detection in images. It calculates the Hessian matrix at each interest point and

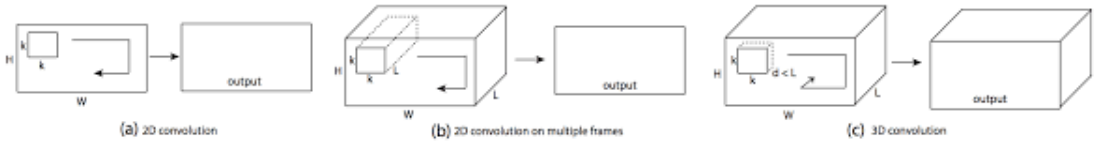

Figure 2.10 : 2D vs 3D convnet models (reprinted from [Tran *et al.* [2015]], © IEEE, 2015)

uses the determinant of the Hessian matrix for point localization and scale selection. The detected points are denser than the Harris3D detector but not as dense as that from the Gabor detector. Many such detectors are developed for human action representation using different filters.

Action videos are represented using different encoding techniques over the detected points. We discuss two of the most famous encodings used for action recognition. The most popular is the bag-of-features encoding. This encoding technique computes a spatial histogram of local feature occurrences in a video sequence. The bag-of-features quantizes local features extracted from the training videos to a visual vocabulary and represents a video using histogram of these quantized features. This technique has been used in many works on action recognition [Schuldt *et al.* [2004a]; Dollár *et al.* [2005]; Niebles *et al.* [2008a]]. Fisher vector encoding (Fisher vectors) [Oneata *et al.* [2013]] is another popularly used encoding technique that represents features as differences between features and visual words instead of directly mapping to the visual words. The clustering of features is done with Gaussian Mixture Model clustering. These encoded representations of videos are directly fed into the classifier to predict the action label of the video.

The local features discussed in this section are invariant to illumination changes, scale, rotation, etc. Further they are generalized and can be applied to model any type of activity. These features have shown high action recognition accuracy. They are well suited for activity level monitoring where it is important to check that the actions are in a right order, and to analyse the timings of the sub-actions. However, they are not suitable for monitoring applications where assessment related to correctness of postures is needed.

2.2.4 Features from Deep Learning Paradigm

Till now we saw that the traditional spatio-temporal approaches of human action recognition involved three phases - feature extraction (using STIP/Gabor like features), feature encoding using bag-of-words or fisher vectors and lastly a classifier like SVM to classify into known action classes. Convolutional Neural Networks, replace these three stages with a single neural network that is trained end-to-end from image pixel values to classifier outputs.

Compared to image data domains, there is relatively less literature on application of CNNs to video classification. This is probably attributed to lack of large-scale video classification benchmarks. Despite these limitations, some extensions of CNNs for the video domain have been introduced.

The spatio-temporal information in videos can be encoded by training a 3D CNN where the convolution kernels in CNN are extended from 2D to 3D. Such networks leverage both visual appearance of the video frames and temporal information in the sequence of frames.

C3D [Tran *et al.* [2015]]. a deep 3-dimensional convolution network, uses a homogeneous network architecture with $3 \times 3 \times 3$ convolution and pooling filters that transfer spatio-temporal information throughout the network (Figure 2.10). While C3D network has shown a remarkable classification performance, training the network from scratch for the task of assessment is computationally complex and the model size is another bottleneck of the architecture. For C3D

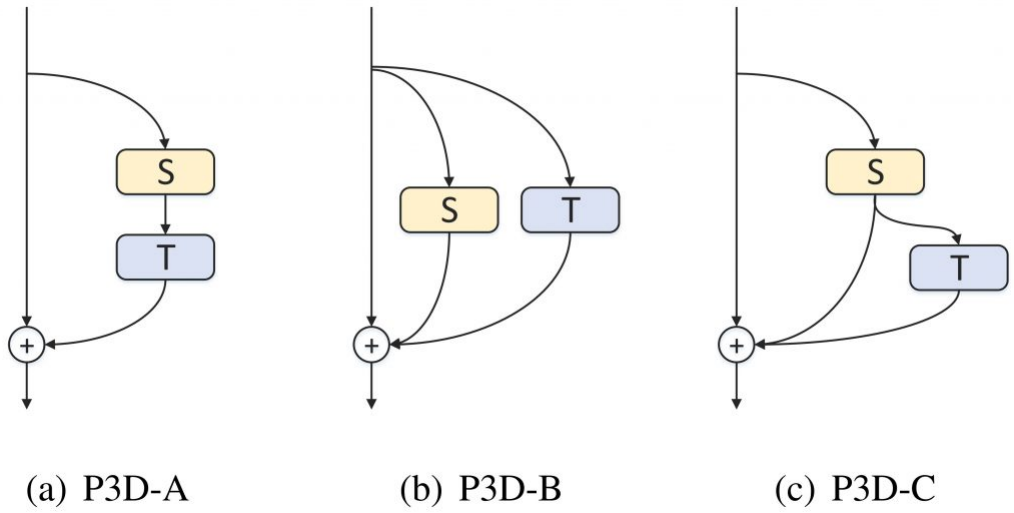

Figure 2.11 : 3D pseudo resnet : an alternate to 3D convolution with separate spatial and temporal filters (reprinted from [Qiu *et al.* [2017]], © IEEE, 2017)

networks, the model size reaches $321MB$.

The above limitations are mitigated by Qiu *et al.* [2017] who devised Pseudo-3D Resnet(P3D Resnet) architecture that contains a family of bottleneck building blocks which leverage both spatial and temporal convolutional filters. The key component in each block is a combination of one $1 \times 3 \times 3$ convolutional layer and one layer of $3 \times 1 \times 1$ convolutions in a parallel or cascaded fashion, which replaces $3 \times 3 \times 3$ convolutional layer (Figure 2.11). Thus, the model size is significantly reduced and the advantages of pre-learned 2D CNN in image domain could also be utilized.

Similarly there are a few more deep networks that are designed to generate spatio-temporal representation of the videos. Deep networks based representations have been used in the past assessment works [Parmar and Tran Morris [2017]; Xiang *et al.* [2018]; Doughty *et al.* [2018a,b]]. These are suitable for cases of actions with objects and events like splash. However they have a limitation that deep learning based features cannot provide appropriate pose-level feedback.

Further, the cited assessment works utilize pre-trained models to derive deep features. These pre-trained models have been trained for the task of classification. Such features cannot be learned for the assessment tasks due to lack of videos in assessment related datasets.

Thus for tasks that require high precision assessment accuracy, marker-based systems are the most preferable options. Deep pose estimation libraries for RGB camera frames can be used in other tasks where posture assessments is important but little errors due to pose estimation errors are not a serious concern. For tasks that involve objects and splash-like outcomes deep convolution networks are a good choice.

2.3 HUMAN ACTION RECOGNITION

Videos are composed of multiple XY images stacked sequentially in time T. Based on how we treat T, we broadly divide human action recognition approaches into two categories : space-time approaches and sequential approaches. Space-time approaches consider video as a 3-dimensional XYT volume while sequential approaches interpret it as a sequence of observations, i.e. XY is treated

differently from T. We review some of the works under these categories.

2.3.1 Space-Time approaches

As seen in the previous section, three common ways of representing actions are as space-time volumes or using space-time interest point or using deep spatio temporal features. All these techniques consider video as a 3-dimensional XYT volume.

Space-time volume

The space-time volume based approaches either concatenate the 2-D images with respect to time as in motion energy (MEI) and motion history (MHI) images [Bobick and Davis [2001]] which have used template based matching techniques for recognizing human action. Another representation by Blank *et al.* [2005] have stacked the silhouettes of the humans to generate a space-time volume. Hierarchical similarity measurements over these volumes are used for action recognition.

Both of these methods are suitable only for actions performed in uncluttered environments where the human silhouette extraction is easier and clean. For cluttered performance the background subtraction is difficult. Further it was also seen that the approach taken by Bobick and Davis [2001] fails for complex activities due to the overwriting of motion history.

Space-time interest points

While discussing spatio-temporal features, we represented video sequences by encoding the extracted interest points. This encoded representation is used for action classification. We briefly present two classification techniques here.

Instance based classifier is one category of classifiers. k-Nearest Neighbor algorithm is a part of this classifier category. It locates the k instances nearest to the given query instance and assigns the label that occurs most frequently in these neighboring instances. Efros *et al.* [2003b] uses k-Nearest Neighbor to classify actions using optical flow features. The classifier needs to store all the instances and is sensitive to the choice of the similarity function used. This limits its use in the action recognition field.

Support Vector Machine (SVM) is another category of classifiers. It maximizes the distance between a hyperplane that separates two classes of data and instances on either side of it. They are capable of performing both linear separation and non-linear separation using a kernel function. They reach a global minimum of the objective function, unlike neural networks. [[Dollár *et al.*, 2005; Laptev, 2005a; Laptev *et al.*, 2008a; Niebles *et al.*, 2008a]] are some of the works in this category. Support vector machines are seen to give best action classification results. However, their performance relies highly on the type of kernel function.

Deep Features

As discussed in the previous section, deep learning models replace three stages of human action recognition problem - feature extraction, encoding, and classification, with a single neural network that is trained end-to-end from image pixel values to classifier outputs. Deep features learned as fully connected layers are used as inputs for Support Vector Machines to give classification results. Psuedo-3D models [Qiu *et al.* [2017]], C3D features [Tran *et al.* [2015]], Resnet-152 features [He *et al.* [2016]] have been along with linear SVMs to classify human actions. Qiu *et al.* [2017] tested these networks for multiple human action recognition datasets like UCF101 [Soomro *et al.* [2012]], ActivityNet [Caba Heilbron *et al.* [2015]], ASLAN [Kliper-Gross *et al.* [2012]], YUPENN [Derpanis *et al.* [2012]]. The deep models are seen to outperform Space time Interest Point (STIP) features.

2.3.2 Sequential Modeling

Sequential modeling approaches explicitly model the variations of actions with time by considering frames in the chronological order. Features are extracted at all time instances for describing the motion of the person in each frame. Action classes are modeled over the sequence of feature vectors extracted from training videos. Sequential approaches analyze a test sequence by comparing the sequence of its feature vector with the vector representations of the input classes.

Sequential approaches are further categorized into three categories using (i) dynamic time warping (ii) using state space models and (iii) deep sequential networks. Dynamic time warping compares the input video directly with the training videos to find which of the action class gives the highest match score. State space models construct a model which is trained to generate sequences of feature vectors corresponding to the action. Actions are recognized based on the maximum likelihood values from the models. Recurrent Neural Networks are popularly used deep sequential networks that build temporal connections over high level features like C3D, pose features etc.

Action recognition using Dynamic Time Warping

Dynamic time warping is a distance measure between two input sequences that can have different temporal lengths. The two sequences have to be segmented in time as this approach needs segmented boundaries between which the cost is to be evaluated. The alignment of the two sequences is done using Dynamic Programming. Veeraraghavan *et al.* [2006] have applied DTW for human action recognition and recognized actions like picking an object, pushing, waving and throwing.

State space models

State space models group features into similar configurations, i.e. states, and learn temporal transition functions between these states. Such models fall into the class of probabilistic graphical models. These models can be generative or discriminative. Generative models learn a joint distribution over both the observations and the action labels and thus learn to model an action class with all its variations. In contrast, discriminative models learn the conditional probability of the observations given a class label. They do not model a class but the difference between the classes.

Among the generative models used for action recognition, the most prominent is certainly the hidden Markov model (HMM) [Rabiner [1989]] which gained its importance because of its great success in the speech and natural language processing community. The first work on action recognition using HMMs is by Yamato *et al.* [1992a], where a discrete HMM is used to represent sequences over a set of vector quantized mesh features of tennis footage. Starner and Pentland [1997] use a continuous HMM for recognition of American sign language. Brand *et al.* [1997] learn coupled HMMs to model interactions between several state variables, e.g. interactions between left and right hand motions. HMM-MIO [Concha *et al.* [2011a]] have proposed an extended version of HMM to model irregular observation like STIP features.

A tractable generative model has its limitation in the form of the independence assumption between the observations in time for computing the joint probability of states and observed features, whereas a more general discriminative model may better predict the conditional probabilities of states given the observed features. Since the discriminative models are trained to differentiate between classes rather than learning the class specific distributions, they are well suited for identifying classes where HMMs fail to distinguish. As a result, several authors have investigated the use of discriminative models for action classifications.

Conditional random fields (CRF) [Sutton and McCallum [2006]] are widely used discriminative model used in the field of action recognition. Mendoza and De La Blanca [2008] have achieved better results for action recognition using CRF over shape features as compared to the HMMs. Natarajan and

Nevatia [2008] use a two-layer model where the top layer models the actions and viewpoints and the lower level CRFs encode the action and viewpoint specific postures.

Deep Sequential Models

Different types of sequential deep learning models exists in literature : Recurrent Neural Networks (RNN), Long-Short Term Memory (LSTM) and Gated Recurrent Units (GRU). The structure of these networks allow them to solve problems involving time series. The benefit of using these networks for sequence classification is that these networks can learn from the raw time series data directly and do not require domain expertise of manually engineering the input features as in the case of state space models. These models have been readily used for the task of human action recognition, both for applications involving acquisition from Kinect cameras [Veeriah *et al.* [2015]; Du *et al.* [2015]; Zhu *et al.* [2016]] and RGB cameras [Baccouche *et al.* [2010]; Grushin *et al.* [2013]; Shi *et al.* [2017]].

2.4 COMPLEX ACTION SEGMENTATION

Till now we have been discussing the representation and recognition techniques for atomic actions. Long-term actions or complex actions are sequences of atomic actions and the techniques discussed so far are not directly applicable to complex actions. Such actions require an additional segmentation and localization step such that in addition to action recognition, the start and end frames of the action need to be identified too.

It is important to note that for the action segmentation techniques to be suitable for the task of long-term action assessment, they should have an additional capability of identifying anomalous or unseen actions.

Early works towards human action segmentation treated segmentation and action recognition as two independent steps. One such strategy, is to use a sliding window-based segmentation followed by approaches like Bag-of-Words [Duchenne *et al.* [2009a]] or unsupervised human action recognition techniques like probabilistic latent semantic analysis (pLSA) and Latent Dirichlet Allocation (LDA) [Niebles *et al.* [2008a]] that perform action recognition over pre-segmented videos. Such techniques result in segmentation errors contributing to classification errors. Further, these techniques could identify known classes of human actions only.

This encouraged techniques where top-down action recognition assisted bottom-up segmentation, giving us principled approach to human action segmentation and recognition. The joint action segmentation and recognition are broadly of two types: *Generative models* and *Discriminative models*.

Generative models: Hidden Markov Model is the most extensively used technique in this category [Borzeshi *et al.* [2013a]]. Generative models rely on computation of joint probability of states and observed features. It gives a posterior probability of assigning labels to all observations. In case an unknown action appears, the posterior of assigning a known class label would be low and the segment would be identified as an unknown class.

Discriminative models may better predict the conditional probability of the states given the observed features. Conditional Random Fields and Support Vector Machines are commonly used techniques in this category. Discriminative models cannot easily handle unseen observations. They classify such observations to one of the known classes over which they have been trained. However there are works that mitigate this limitation and are able to segment videos in the presence of unseen observations. Shi *et al.* [2008] proposed a semi-Markov model-based framework where emphasis was on properties of both the individual segments and adjacent action segments, unlike HMM and CRF

that consider statistical dependencies over adjacent frames. Hoai *et al.* [2011] proposed a joint action segmentation and recognition technique based on multi-class SVM and dynamic programming approach to find the candidate segments of each class by maximizing the confidence of segment assignment. Recently, a discriminative Hough Transform-based approach [Kosmopoulos *et al.* [2011, 2016]] has been proposed, where the putative segments are detected by collecting votes generated by action primitives.

Common to all the works discussed so far is that they all use training videos to develop a model which is then applied over the test videos to find segmentation and recognition results. These methods require a lot of training data. Annotating such videos for supervised training is quite labor intensive. However, there are scenarios where the collection of such a huge number of videos is difficult, eg. healthcare and surveillance. To deal with such scenarios, unsupervised clustering-based methods have been developed that parse actions into motion primitives. There have been many promising techniques developed in this direction.

Zhou *et al.* [2008] and Zhou *et al.* [2013] proposed Aligned Cluster Analysis (ACA) and its improved version - Hierarchical Aligned Cluster Analysis (HACA), a novel kernel for time series alignment. Upon specifying the number of clusters, as well as the minimum and maximum lengths of the action segments, these techniques solve a dynamic program over the entire stream to find putative action class segments. Krüger *et al.* [2017] proposed an efficient motion segmentation approach, in which a novel feature bundling method is used to generate compact and robust motion representations. A generalized search radius is introduced so that the number of clusters is not needed as input. This technique works efficiently for periodic motion only. These unsupervised clustering techniques have known to give promising results for human action segmentation and recognition. However they have no added machinery to discriminate between the known and unobserved features like in the previously discussed works [Shi *et al.* [2008]; Hoai *et al.* [2011]; Kosmopoulos *et al.* [2011, 2016]]. i.e. they work under the assumption that all frames in the given sequence are a part of one of the known classes. This is again not suitable for the task of segmenting complex actions with anomalous sub-segments.

A segmentation technique similar to Aligned Cluster Analysis techniques [Zhou *et al.* [2008, 2013]] that does not require an explicit training phase and has an added capability of unseen action segment detection, as in [Shi *et al.* [2008]; Hoai *et al.* [2011]; Kosmopoulos *et al.* [2011, 2016]], remains an open challenge. We propose a technique in this direction and discuss in Chapter 7.

In this chapter, we have reviewed past literature on human action analysis that includes human action representation, recognition, and complex action segmentation. The techniques discussed in this Chapter are relevant and used in the task of human action assessment. We discuss the literature in the area of human action assessment in the next Chapter.

...