Chapter 1 INTRODUCTION

This chapter presents our motivation behind the present work, an introduction to the problems discussed and objectives of this thesis. Finally, the chapter concludes with a list of contributions of this thesis in the field of human activity recognition.

1.1. Background

Video-based human activity recognition has fascinated researchers of computer vision community due to its critical challenges and a wide variety of applications in the surveillance domain. Thus, the development of techniques related to human activity recognition has accelerated. There is now a trend towards implementing deep learning based activity recognition systems because of performance improvement and automatic feature learning capabilities.

The overall task of human activity recognition consists of two main steps, feature representation followed by the classification, both of which can be accomplished in different ways. Despite tremendous progress in the annotation of video streams with corresponding activity classes, activity recognition methods may still misclassify videos due to various difficulties as identified by Ronald Poppe [1]. Earlier work on activity recognition mainly involves global and local feature representations. Global features capture holistic information of the whole scene comprises of geometric structure, motion and appearance. Common examples are Motion energy image (MEI), Motion history image (MHI) introduced in Bobick et al. [2]. These global representations are variant to noise and cluttered background. Local features such as spatio-temporal interest points (Ivan Laptev [3]) describe human motion in space–time regions. The main advantage of spatio-temporal interest points is robustness to translation and appearance variation. But they capture only short-term duration information; trajectories are useful at capturing long-term information. Despite the great success achieved by local and global feature methods, handcrafted features are quite labour intensive and demands for the domain knowledge. Therefore, feature extraction and classification using deep learning is receiving increasing attention. Deep learning is a particular kind of machine learning that achieves great power and flexibility by learning to represent the world as a nested hierarchy of concepts, with each concept defined in relation to simpler concepts, and more abstract representations computed in terms of less abstract ones. Deep learning models are thus capable of learning a hierarchy of action features by building high-level features from low-level ones, automating the process of feature extraction. Thus deep learning based solutions do not require having domain knowledge and a heavy burden of feature engineering.

1.2. Motivation

Recent advances in digital imaging technology, computational speed, storage capacity, continuous increase in computational capabilities of hardware at low cost, and networking have made it possible to capture, manipulate, store, and transmit images and videos at interactive speed with equipment available at home or business. As a result, images and videos are now becoming an integral part of our day-to-day life, and are being used for entertainment, education, medicine, security, science and a number of other applications. Some of the example applications of intelligents surveillance system has been shown in figure 1.1.

Public area surveillance

Traffic monitoring

Intelligent environments

Sports analysis

Figure 1.1 Applications of video surveillance system

Computer vision aims to provide description, understanding and interpretation of a scene by extracting the image features. There are several computer vision applications, which are used in a multitude of problems. For example, medical imaging, tumor detection, measurement of size and shape of internal organs etc. In forensics, it is desired that the identity of a person would be ensured. In intelligent video surveillance, it is highly desirable that an intruder can be tracked and recognized. There are other important applications too, like industrial automation, radar imaging, remote sensing, robotics, etc. Now-a-days research activities are getting focused on the development of an intelligent video surveillance system with advances in computer vision methodology and processing capabilities. After terrorist attacks on 9/11 in USA, and that on 26/11 in Mumbai, India, a strong need for improvement in existing video surveillance capabilities was realized by the security establishments across the globe to prevent any such further terrorist attacks. Therefore, challenges and problems associated with the development of intelligent video surveillance system need to be addressed. These challenges and problems motivated me to design and develop a framework for human activity recognition with automatic feature extraction and classification capabilities.

1.3. Problem Statement

Human Activity Recognition in a video surveillance system under complex environment a challenging task. With a large annotated activity dataset, we can train deep neural network models, which can further recognize actions in the new unseen videos.

"Study of Conventional and Deep Learning models for human activity recognition and Implementation of frameworks to identify human activities in a video surveillance system by using convention machine learning and deep learning approaches".

The prime objective of the research is to develop an automatic human activity recognition model that can better exploit available data with minimal pre-processing, hence can cope up with above said challenges. In order to achieve this goal, a number of tasks must be accomplished.

1.4. Thesis Objectives

The objective of this thesis is to apply deep learning and conventional learning models on videos datasets and propose different machine learning approaches for human activity recognition with better accuracy for the complex state of the art datasets. Firstly, a HAR model using HMM has been proposed and tested on controlled/ lab environment datasets like KTH, i3DPost multi-view dataset, and MSR view-point action dataset with a very good accuracy. But, when the model was tested for modern complex datasets, accuracy was abysmal. Subsequently, in this thesis, several models using deep learning frameworks has been proposed and evaluated over the modern state of the art datasets (NTU-RGB+D, MSRAction3D and MSRDailyActivity3D, UCF-101, MSR daily Activity, UTD MHAD and CAD 60) with accuracy comparable to state of the art models proposed in the literature.

The objectives of this thesis are as follows:

- An extensive study of the existing literature on experiments and research performed under Human Activity Recognition using conventional as well as deep learning approaches to identify research gaps.
- An extensive study of literature on the evolution of modern datasets for human activity recognition
- Proposing a new handcrafted feature based machine-learning approach for human activity recognition, its implementation and performance evaluation.
- Proposing deep learning based approaches for human activity recognition, their implementation and performance evaluation to address the limitation of the existing methods.

1.5. Contribution to the Thesis

The major contribution to the thesis may be summarized as follows-

(i) Performed extensive study of the existing literature on experiments and research under Human Activity Recognition using conventional as well as deep learning approaches to identify research gaps.

(ii) Performed extensive study of literature on the evolution of modern datasets for human activity recognition

(iii) Proposed a conventional machine learning based human activity recognition framework using HMM and multiple features.

(iv) Proposed and evaluate the performance of enlarged temporal convolution network on depth sequences received from RGB-D Sensors. Further, analyzing the performance of the model for varying temporal dimension.

(v) Proposed an encoder-decoder based framework using ResNet CRNN. Evaluation of the proposed model by training the network from scratch and by using pre-trained ResNet.

(vi) Proposed a two-stream model using spatial and spatio-temporal streams for activity recognition from videos. Implemented the proposed model and fine-tune pre-trained model for proposed solution. Further, analyzed the impact of network depth on the performance of the proposed model

(vii) Proposed a four-stream model using depth and RGB images generated from RGB-D Sensors. Dynamic images from RGB input using rank pooling were generated and used as the first stream. Further, projection of depth motion maps in three different directions, namely front, side and top were generated and used as three different streams to the model. Using these 4 input streams, individual pre-trained CNNs were trained. Finally, for the outputs received from each stream, decision fusion was done using a weighted product model to obtain the classification scores.

1.6. Thesis Organization

The overall thesis is organized into seven chapters as follows

Chapter 1 presents a brief introduction of the problems addressed in this thesis, followed by the objectives of the thesis. Finally, the chapter concludes with a brief account on contributions of this thesis in the field of image/video processing.

Chapter 2 discusses the theoretical background of a video surveillance system. In this chapter, we have also given an overview of deep learning and machine learning approaches. Further, in this chapter a literature survey of prominent approaches for HAR using conventional and deep learning approaches are given. Furthermore, it presents a detailed survey along with the evolution of modern datasets for HAR.

Chapter 3 presents a HAR method based on the conventional machine learning approach. The complete framework comprises of three steps. In the first step, we perform preprocessing and background subtraction. Secondly, feature computation is done. In the third and last step, we have used HMM for activity modeling and classification.

Chapter 4 presents a Deep Learning based HAR approach utilizing depth map sequences as input. The proposed approach uses an enlarged temporal dimension of depth map sequences as input for training the deep neural architecture. The impact of enlarged larger temporal and spatial resolution has been evaluated on three HAR depth datasets, namely NTU-RGB +D, MSRAction3D and MSRDailyActivity3D. From the experimental results, it can be observed that the result obtained is comparable to state of the art models proposed in the literature.

Chapter 5 of the thesis presents HAR models utilizing deep residual networks. The first model is dual stream model using residual-CNN of two streams, namely spatial and spatio-temporal. While the other discussed model is encoder-decoder based model using CRNN, which is a combination of CNN as encoder and RNN as a decoder.

Chapter 6 presents a HAR model by a combination of different modalities from RGB-D sensor. In this work, dynamic images trained on pre-trained VGG-F network and depth motion maps for different views such as top, side and front separately trained on pre-trained VGG-F network are combined. We have tested the model on most promising datasets such as MSR Daily Activity, UTD MHAD and CAD 60 and achieved state-of-the-art results. In addition, we have compared our results for different datasets and found that the proposed method outperforms most of the available methods.

Chapter 7 concludes the thesis and summarizes the main findings of the work done. This chapter also proposes some possible future perspectives of the thesis.