Review Of Improving Healthcare Services Through Human Activity Recognition

Gangireddy Prabhakar Reddy, Dr. M Kalaiselvi Geetha

Abstract: Healthcare industry is the one which renders valuable services to humans. With technological innovations, the quality of healthcare services is increased substantially. With the advancements in image processing and deep learning techniques, new avenues in healthcare domain are being witnessed. Human actions with respect to healthcare services exhibit different viewpoints that can help in understanding the status of health of patients under observation and elders under constant care. Traditional approaches for monitoring human actions are prone to errors and time consuming which lead to deteriorated health conditions and delay in seeking healthcare services. With an automated process driven by technology, humans who are in need can be monitored and observed remotely. This sort of research comes under Human Action Recognition (HAR). It is essential to capture spatio temporal features in order to capture live information pertaining to people under observation. Machine learning with its advancements like deep learning is part of Artificial Intelligence (AI) which provides actionable knowledge by processing human actions in real time. This approach has plenty of advantages including saving time, effort, life and money. Due to its indispensable need, in this paper, we study the state of the art and provide useful insights and research gaps if any in order to improve the research further in the area of HAR.

Keywords – Machine learning, image processing, AI, human activity recognition, healthcare

1. INTRODUCTION

As the saying goes “Health is Wealth”, it truly reveals the need for continuous improvements in healthcare industry. In spite of novel technologies and innovations, healthcare industry still needs optimizations in the area of monitoring humans remotely. Those who need close observation, unlike a traditional system, it is essential to have application of technologies like Internet of Things (IoT) and machine learning (part of AI) for real time monitoring of patients or elderly people whose activities have great meaning to serve them in healthcare domain [1]. The area of research is known as HAR and its process is illustrated in Figure 1.

Figure 1: Typical procedure for HAR

As presented in Figure 1, the human gestures are captured live using a camera or any such device for surveillance. Then the data is given to a remote server for machine learning. The machine learning or deep learning or AI results in actionable knowledge that can help in making well informed decisions. Such knowledge is also known as Business Intelligence (BI). Many machine learning techniques came into existence. Sum of them are related to supervised learning like SVM [1] while others are related deep learning such CNN [2]. Many deep learning techniques [3, 4, 5, 6, 7, 9] are used for different purposes. However, in the area of remote human action recognition, there is room for further research. The contribution of this paper is to introduce the concept of HAR and provides some useful literature insights besides providing research gaps. The remainder of the paper is structured as follows. Section 2 presents the review of literature and its various aspects. Section 3 presents summary of findings and section 4 concludes the paper and provide possible scope for future research.

2. DIFFERENT HAR APPROACHES

HAR has been an active area of research for some years. The research focused with traditional (machine learning) methods and also deep learning methods. This section provides insights on the same. HAR involves an important component known as capturing temporal structure from the given video clip. Gaidon et al. [1] studied the concept of actoms for characterizing human activities that are observed from videos. It makes use of a sequence of representations in the form of histograms to do so. It makes use of an actom sequence model which is non parametric. Based on this action model, it will recognize human actions. When the actions are very complex, they need different approach. Niebles et al. [2] proposed a method known as SVM. It is named as latent SVM that is capable of discriminating between motion segments that are subjected to temporal decomposition. The algorithm proposed is on the base of latent temporal decomposition which is described in [3]. It also depends on the concept of segmental grammar model explored in [4]. Human appearance over time changes based on the actions. Features that are temporally captured can help to understand the changes. Fernando [5] proposed a ranking function that can learn evolution of the appearance in terms of changes in the video captured. It is based on the temporal features and the parameters associated with the same. When it comes to detecting something in an image or video, it is important to capture features in the form of descriptors. Of late, the concept of 3D is being used in the capturing of descriptors. For instance, Scovanner et al. [6] proposed a novel descriptor based on SIFT known as SIFT-3D. It is meant for representing multiple aspects such as video and bag of words. Such descriptor can find spatio-temporal features more effectively. Similarly, the existing
HOG descriptor is enhanced in [7] to have a novel feature descriptor known as HOG-3D. It makes use of histograms that are captured from 3D spatio-temporal gradients that are derived from video sequences in order to characterize human actions. In addition to these descriptors based on 3D, recently, many other descriptors came into existence. They are important for HAR study. Novel descriptors known as high-level descriptors and mid-level descriptors are proposed to lead to formation of an action bank. This action bank is then used for detection of various human actions from different viewpoints like viewpoint space and semantic space in order to have high level of discriminative power. This is achieved in [8] with the machine learning method known as SVM. All human actions are related to human poses either directly or indirectly. Therefore, it is essential to have human pose related knowhow for detection of actions. Keeping this in mind, the concept of dynamic-poselets is introduced in [9] to have a learning model known as relational model to characterize and decompose human actions with the help of key poses and action parts. When it comes to mid-level actions, Zhu et al. [10] focused on the learning with multiple instances and features covering max-margin and multi-channel optimizations. Such learned model can have good information that is scalable, discriminative, informative besides compact. Apart from the aforementioned techniques, deep learning related research focused on the HAR. For instance, CNN is one of the AI techniques that exhibited very good performance for HAR. As studied in [11], deep ConvNets are used widely in order to have CNNs that make use of video inputs and produce outputs related to human actions. They are also found to be scalable in order to support large databases of human activity related images or videos. They use temporal domain to improve the overall performance. Simonyan and Zisserman [12] developed a ConvNet based on the two-stream approach that differentiates temporal features from spatial counterparts. Thus it could exploit dense optical flow that was able to enhance the overall performance even in spite of limited training data available. A novel descriptor based on deep learning is explored in [13]. It is known as trajectory pooled deep convolutional descriptor. It was used to improve performance of the two-stream ConvNet introduced in [12]. This descriptor is proved to be good as it can combine learning features for leveraging learning process. Further improvement is made to the two-stream ConvNet approach in [14]. It involves the concept of long range HAR. Towards this end, Donahue et al. [15] proposed a R-CNN that makes use of LSTM in order to achieve extraction of long-range temporal data. As mentioned earlier 3D concept is used widely. For instance, Tran et al. [16] focused on the usage of CS3 based algorithm that is effective in extracting spatial-temporal features with the help of 3D-CNNs that are described in [17]. Another such research is carried out in [18] where spatial-temporal CNN is used with 3D convolutional kernels decomposed. The 3D CNN with kernel is factorized as if it is 1D on top of 2D kernels in spatial domain. There are videos that are untrimmed and that are common in usage in applications. In such videos, HAR is not easier and that needs action localization. In such cases, a deep segmental work proposed in [19] is used. As shown in [19], the deep segmental network is used to divide the given video into multiple clips. Without this, it is not easier to have segmentation and action localization in untrimmed videos. In the untrimmed videos, Wang et al. [20] proposed a model called attention model in order to achieve HAR. The model has different modules like selection module and classification module. The time taken for different actions is understood from the model. The proposals are used for both temporal actions and spatial regions for localization of actions that are surfacing in untrimmed videos.

3. SUMMARY OF FINDINGS

As found in the literature there are many techniques used for human action recognition directly or indirectly. Localization of actions [1], representing decomposable motion segments [2], complex human activity classification [3], parsing human action videos [4], video evolution modelling [5] are related to HAR. In [6]-[10] spatio temporal features, SIFT descriptor, spatio temporal descriptor, activity representation, notion of dynamic poselets are the techniques used. CNN is used in [11], [12], [13], [14] and [15]. Concept of 3D CNN is used in [16] and [17] while spatio temporal CNN is used in [18]. Deep action recognition is studied in [19]. Supervised learning method is employed in [20] for HAR procedure. R-CNN [21], fast temporal activity proposals [22], multi-state CNNs [23], spatio temporal approach for action recognition [24] and RNN [25] are other important techniques in deep learning used for HAR. Important insights found in the literature are as follows. Human action recognition needs image processing techniques and feature selection techniques. It is to be understood in both temporal and spatial domains. SIFT, SURF, STIP and HOG are important for extracting features from images. The concept of poselets and dense trajectories can be employed. CNN, RNN/R-CNN deep learning methods could improve the performance of HAR. LSTM is one of the deep learning techniques. 3D-CNN is also to be given importance for to capture motion temporally for effectiveness. There is novel descriptor related to deep learning known as trajectory pooled deep convolutional descriptor.

4. CONCLUSION AND FUTURE WORK

In this paper, the concept of HAR is introduced besides reviewing literature on the same. HAR has many applications in real time. Recognition of abnormal activities with respect to anti-social elements is an important area for such applications. Another area which is essential for human health is the healthcare domain. In this paper, the research is carried out using the latter area. Literature revealed that different machine learning techniques like SVM are widely used for image mining. Feature extraction from images such as SURF, SIFT and HOG are used for human action recognition. Deep learning methods like CNN and RNN are used for the same. Techniques like LSTM are used to capture long range temporal data. Deep segmental networks and advancements in deep learning are employed to recognize human actions. From the review of literature, it is found that there is need for further research to enhance the state of the art. We extend this research by exploring...
deep learning further with novelties like transfer learning and image descriptors to have more accurate and faster (near real time) recognition of human actions for the purpose of improving healthcare services.

References


