Leveraging CNN and Transfer Learning for Vision-based Human Activity Recognition

Samundra Deep  
Department of Computing  
Macquarie University  
Sydney, Australia  
Email: samundra.deep@hdr.mq.edu.au

Xi Zheng  
Department of Computing  
Macquarie University  
Sydney, Australia  
Email: james.zheng@mq.edu.au

Abstract—With the advent of the Internet of Things (IoT), there have been significant advancements in the area of human activity recognition (HAR) in recent years. HAR is applicable to wider application such as elderly care, anomalous behaviour detection and surveillance system. Several machine learning algorithms have been employed to predict the activities performed by the human in an environment. However, traditional machine learning approaches have been outperformed by feature engineering methods which can select an optimal set of features. On the contrary, it is known that deep learning models such as Convolutional Neural Networks (CNN) can extract features and reduce the computational cost automatically. In this paper, we use CNN model to predict human activities from Wiezmann Dataset. Specifically, we employ transfer learning to get deep image features and trained machine learning classifiers. Our experimental results showed the accuracy of 96.95% using VGG-16. Our experimental results also confirmed the high performance of VGG-16 as compared to rest of the applied CNN models.

Index Terms—Activity recognition, deep learning, convolutional neural network.

I. INTRODUCTION

Human activity recognition (HAR) is an active research area because of its applications in elderly care, automated homes and surveillance system. Several studies has been done on human activity recognition in the past. Some of the existing work are either wearable based [1] or non-wearable based [2] [3]. Wearable based HAR system make use of wearable sensors that are attached on the human body. Wearable based HAR system are intrusive in nature. Non-wearable based HAR system do not require any sensors to attach on the human or to carry any device for activity recognition. Non-wearable based approach can be further categorised into sensor based [2] and vision-based HAR systems [3]. Sensor based technology use RF signals from sensors, such as RFID, PIR sensors and Wi-Fi signals to detect human activities. Vision based technology use videos, image frames from depth cameras or IR cameras to classify human activities. Sensor based HAR system are non-intrusive in nature but may not provide high accuracy. Therefore, vision-based human activity recognition system has gained significant interest in the present time. Recognising human activities from the streaming video is challenging.

Video-based human activity recognition can be categorised as marker-based and vision-based according to motion features [4]. Marker-based method make use of optic wearable marker-based motion capture (MoCap) framework. It can accurately capture complex human motions but this approach has some disadvantages. It require the optical sensors to be attached on the human and also demand the need of multiple camera settings. Whereas, the vision based method make use of RGB or depth image. It does not require the user to carry any devices or to attach any sensors on the human. Therefore, this methodology is getting more consideration nowadays, consequently making the HAR framework simple and easy to be deployed in many applications.

Most of the vision-based HAR systems proposed in the literature used traditional machine learning algorithms for activity recognition. However, traditional machine learning methods have been outperformed by deep learning methods in recent time [5]. The most common type of deep learning method is Convolutional Neural Network (CNN). CNN are largely applied in areas related to computer vision. It consists series of convolution layers through which images are passed for processing. In this paper, we use CNN to recognise human activities from Wiezmann Dataset. We first extracted the frames for each activities from the videos. Specifically, we use transfer learning to get deep image features and trained machine learning classifiers. We applied 3 different CNN models to classify activities and compared our results with the existing works on the same dataset. In summary, the main contributions of our work are as follows:

1) We applied three different CNN models to classify human recognition activities and we showed the accuracy of 96.95% using VGG-16.

2) We used transfer learning to leverage the knowledge gained from large-scale dataset such as ImageNet [6] to the human activity recognition dataset.

The rest of the paper is as follows: Section II provide an overview of the related work in video-based HAR systems. We provide an overview of transfer learning in section III. Section IV outline the research methodology, sources of data, research approach and discuss the experimental results. Conclusion and future work are drawn in Section V.
II. RELATED WORK

There have been a lot of research on vision-based human activity recognition in recent years. Most of the studied methods depend on handcrafted feature extraction from the videos/images and employed traditional classifiers for activity recognition. The traditional approaches often achieved optimum results and exhibited high performances. However, traditional methods are not feasible to deploy in real life because handcrafted features are highly dependent on data and are not robust to the environment change.

Hidden Markov Model (HMMs) methods have been largely used as the recognition techniques in the past because of its capability of temporal pattern decoding [7]. However, researchers are more interested in using deep learning techniques because of its ability to automatically extract the features and learn deep pattern structures [5] [7]. Deep learning methods have clearly ruled out traditional classification methods in the domain of computer vision [5] [8]. Deep learning techniques have been largely employed recently in the domain of computer vision and have achieved tremendous results. Therefore, video-based human activity recognition using deep learning models have gained a lot of interest in recent years [5].

Zhu et al. [4] proposed an action classification method by adding a mixed-norm regularization function to a deep LSTM network. One of the most popular deep learning methods in frames/image processing is Convolutional Neural Network (CNN). There have been several works that utilized 2D-CNNs that take advantages of spatial correlation between the video frames and combine the outputs employing different strategies [9]. Many have also used additional input such as optical flow to 2D-CNN to get temporal correlations information [10]. Subsequently, 3D-CNNs [11] were introduced that demonstrated exceptional results in the classification of videos and frames.

Wang et al. [12] applied CNN to RGB and depth frames to automatically extract the features. The obtained features were passed through a fully connected neural network and achieved an improved accuracy. Ji et al. [13] proposed a 3D CNN model which performs 3D convolutions and extract spatial and temporal features by capturing the motion information for activity recognition. Simonyan et al. [8] introduced ConvNet, a two-stream convolution layer architecture that could achieve good results despite of limited training data. Kaire et al. [14] proposed a model that train convnets from RGB-D dataset and combined the softmax scores from depth, motion and skeleton images at the classification level to identify the activities. Karpathy et al. [15] proposed the extension of CNN architecture in the first convolutional layers over a 4D video chunk. Similarly, Tran et al. [16] used a deep 3D CNN architecture (quiet similar to VGGnet [17]) that utilise spatiotemporal convolutions and pooling in all layers to improve the accuracy of the model.

In comparison, we are more interested to explore how transfer learning can be leveraged with CNN models on benchmark dataset to improve classification accuracy.

III. TRANSFER LEARNING

Transfer learning [18] is a method of transferring knowledge that a model has learned from earlier extensive training to the current model. The deep network models can be trained with significantly less data with transfer learning. It has been used to reduce training time and improve accuracy of the model. In this work, we use transfer learning to leverage the knowledge gained from large-scale dataset such as ImageNet. We first extract the frames for each activities from the videos. We use transfer learning to get deep image features and trained machine learning classifiers. For all CNN models, pre-trained weights on ImageNet are used as starting point for transfer learning. ImageNet [6] is a dataset containing 20000 categories of activities. The knowledge is transferred from pre-trained weights on ImageNet to Weizmann dataset, since set of activities recognised in this work fall within the domain of ImageNet. The features are extracted from the penultimate layer of CNNs. The basic idea of transfer learning is as shown in Figure 1.

The main approaches in transfer learning are: (1). to preserve the original pre-trained neural model of large-scale dataset and update weights of the trained model on the target dataset, and (2). use the pre-trained neural model for feature extraction and representation followed by a generic classifier such as Support Vector Machine or Logistic Regression.

IV. IMPLEMENTATION

A. Dataset

In order to evaluate the effectiveness of the models, we do experiments on benchmark activity recognition, namely Weizmann dataset. It consists of 90 low-resolution video sequences showing 9 different people performing 10 activities namely, bend, jack (or jumping-jack), jump (or jump-forward-on-two-legs), pjump (or jump-in-place-on-two-legs), run, side (or gallopsideways), skip, walk, wave1 (waveone-hand), and wave2 (wave-two-hands). We used nine actions (not including pjump jump-in-place-on-two-legs) for our experiment. We first convert all videos into individual frames based on their...
activity. Table I shows the total number of frames per activity based on the extracting frames for all 9 people. The entire dataset is divided into Training (70%), Validation (10%), and Testing (20%).

<table>
<thead>
<tr>
<th>Activity</th>
<th>Number of Frames</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bend</td>
<td>639</td>
</tr>
<tr>
<td>Jack</td>
<td>729</td>
</tr>
<tr>
<td>Jump</td>
<td>538</td>
</tr>
<tr>
<td>Run</td>
<td>346</td>
</tr>
<tr>
<td>Side</td>
<td>444</td>
</tr>
<tr>
<td>Skip</td>
<td>378</td>
</tr>
<tr>
<td>Walk</td>
<td>566</td>
</tr>
<tr>
<td>Wave1</td>
<td>653</td>
</tr>
<tr>
<td>Wave2</td>
<td>624</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>4917</strong></td>
</tr>
</tbody>
</table>

B. Discussion and Results

In order to classify activities, we experiment with 3 different Convolutional Neural Networks (CNN) for activity recognition, namely VGG-16, VGG-19 and Google’s InceptionNet-v3. We used transfer learning to leverage the knowledge gained from large-scale dataset such as ImageNet. Transfer learning technique transfer knowledge from pre-trained model to train a new domain in neural network. We performed experiment on Weizmann dataset using the knowledge learned from pre-trained weights on ImageNet. The features are extracted from the penultimate layers of CNNs. We applied transfer learning on VGG-16 CNN model and achieved accuracy of 96.95%. For VGG-16, image of dimensions 224 × 224 is given as an input and features from fc1 layer are extracted which gives 4096-dimensional vector for each image.

We also applied transfer learning to other CNN models such as VGG-19 and Google’s InceptionNet-v3 to examine the performance of the different CNN models. VGG-19 and Google’s InceptionNet-v3 achieved 96.54% and 95.63% respectively. Experimental results showed that VGG-16 performs better than the rest of the CNN models after all the models have been applied transfer learning. Table II reports accuracy score, precision, recall, and f1-score of the applied CNN models. The confusion matrix of 3 different CNN models are as shown in Figure 2, 3 and 4.

We compared the results achieved from some of the other approaches that do not employed transfer learning on Weizmann dataset with our approach. The experiment results showed that applying transfer learning to the same dataset achieved better recognition scores. The recognition accuracy is improved by 1 − 6% by applying transfer learning. The comparison of results utilising transfer learning on VGG-16 model and other approaches are presented in Table III. The comparison with state-of-the-art approaches is done to explore how effective is transfer learning when leveraged with CNN models for improving recognition scores.
We used CNN models to predict the human activities from Wiezmann Dataset. We experimented with 3 different Convolutional Neural Networks (CNN) for activity recognition. We have employed transfer learning to get the deep image features and trained machine learning classifiers. Our experimental results showed the accuracy of 96.95% using VGG-16 with the implementation of transfer learning. Our experimental results showed that VGG-16 outperformed other CNN models in terms of feature extraction. Our experimental results with transfer learning technique also showed high performance of VGG-16 as compared to state-of-the-art methods.

In future, we aim to extend this study by developing the context-aware recognition system to classify human activities. Also, we will extend our work to recognise complex human activities such as cooking, reading books, and watching TV.

REFERENCES