Human Activity Recognition in Videos

Anurag Sanyal (B14CS007)
Mahak Jain (B14CS042)

I. ABSTRACT

The project deals with the problem of classifying real-world videos by human activity.
Recognizing human actions find applications in a variety of domains including intelligent video surveilance, customer attributes, and shopping behavior analysis.
We would be using 3-dimensional convolutional networks(3D ConvNets) pre-trained on a large scale video dataset for the purpose of feature extraction.This model extracts features from both the spatial and the temporal dimensions by performing 3D convolutions, thereby capturing the motion information encoded in multiple adjacent frames Then we propose to use a simple linear classifier and compare the results with that in paper[2]. We apply the developed models to recognize human actions in the real-world environment , and it achieves superior performance in comparison to baseline methods. After that we would use time sequence models like Hidden Markov Model (HMM) or Recurrent Neural Network for classification.

II. INTRODUCTION

Multimedia on the Internet is growing rapidly resulting in an increasing number of videos resullting in an increasing number of videos being shared every minute.To combat the information explosion it is essential to understand and analyze these videos for various purposes like search, recommendation, ranking etc.
Most current methods build classifiers based on complex handcrafted features computed from the raw inputs. Convolutional neural networks (CNNs) are a type of deep models that can act directly on the raw inputs. In this project, we use a 3D CNN model for action recognition. As a class of deep models for feature construction, CNNs have been primarily applied on 2D images. A simple approach in this direction is to treat video frames as still images and apply CNNs to recognize actions at the individual frame level. Indeed, this approach has been used to analyze the videos of developing embryos [2]. However, such approach does not consider the motion information encoded in multiple contiguous frames.To this end, we propose to perform 3D convolutions in the convolution stages of CNNs to compute features from both spatial and temporal dimensions. The 3D convolution is achieved by convolving a 3D kernel to the cube formed by stacking multiple contiguous frames together. By this construction, the feature maps in the convolution layer is connected to multiple contiguous frames in the previous layer, thereby capturing motion information

The proposed features with a simple linear model outperform the current best methods on 4 different tasks and 6 different benchmarks according to [1].Compared to 2D ConvNet, 3D ConvNet has the ability to model temporal information better owing to 3D convolution and 3D pooling operations. In 3D ConvNets, convolution and pooling operations are performed spatio-temporally while in 2D ConvNets they are done only spatially. Figure 1 illustrates the difference, 2D convolution applied on an image will output an image, 2D convolution applied on multiple images also results in an image. Hence, 2D ConvNets lose temporal information of the input sig- nal right after every convolution operation. Only 3D convolution preserves the temporal information of the input signals resulting in an output volume. The same phenomena is applicable for 2D and 3D polling.

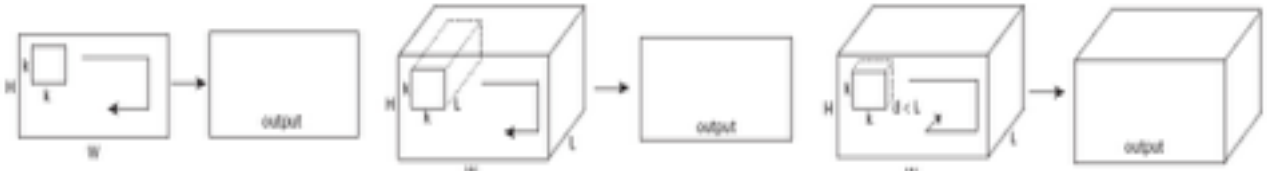

Figure 1 **2D and 3D convolution operations. [1]**

III. CONTRIBUTIONS OF THIS WORK

To effectively incorporate the motion information in video analysis, we propose to perform 3D convolution in the convolutional layers of CNNs so that discriminative features along both the spatial and the temporal dimensions are captured. These 3D feature extractors operate in both the spatial and the temporal dimensions, thus capturing motion information in video streams.By applying multiple distinct convolutional operations at the same location on the input, multiple types of features can be extracted.It has been shown that, when trained with appropriate regularization , CNNs can achieve superior performance on visual object recognition tasks. In addition, CNNs have been shown to be invariant to certain variations such as pose, lighting, and surrounding clutter .

IV. LITERATURE SURVEY

Our model is based on the approach proposed by DuTra et. al. in [1]. in which features learned by C3D are fed to a linear classifier. The results turned out to be outperforming state-of-the-art on 4 different benchmarks and were comparable to the current best methods on the other 2 benchmarks.The model achieved an accuracy of 52.8% on UCF101 with only 10 dimensions.Shuiwang Ji et. al. also devloped a model based on 3D CNN in [2].The devloped model generates multiple channels of information from the input frames, and the final feature representation combines information from all channels and devloped an architecture on TRECVID data set.The 3D CNN model achieves an accuracy of 90.2% as compared with 91.7% achieved by HMAX model. The HMAX model use handcrafted features computed from raw images with 4-fold higher resolution

V. Methodology

We would use pre-trained 3D Convolutional Neural Network for feature extraction.The extracted features will be fed to a linear Support Vector Machine for classification.But, SVM ignores temporal information of data and it is not capable of time sequence modelling.Therefore in later phase, we would be using HMM or RNN as classifier for extracted features.

VI. Work Plan

We have divided the entire project in three phases.

1. Feature Extraction which we expect to complete by 12th of March.
2. SVM (Support Vector Machines) on the features extracted. - To be completed by 20th of March.
3. Use time sequence Model like HMM or RNN on the extracted feature. — To be completed by 12th of April.

a) Distribution of Work:

Phase 1 : Anurag Sanyal

Phase 2: Mahak Jain

While the second phase is going on Anurag will study about RNN.And Mahak will study about HMM during phase 1. After the completion of phase 2 the knowledge and research will be compiled and and an appropriate time-sequence model would be selected .From there onwards both will work simultaneously on devloping the time sequence model as a classifier.

VII. Deliverables

At the end of this project we expect to prepare a model that would be able to predict the label of the activity going on in the video which is given as input.

IX. References.

1. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3D convolutional networks In ICCV, 2015.
2. Ji, W. Xu, M. Yang, and K. Yu. 3D convolutional neural networks for human action recognition.PAMI, 35(1):221–231, 2013.
3. A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In CVPR, 2014.
4. J. Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici. Beyond short snippets: Deep networks for video classification. In CVPR, 2015