International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 4, Issue 9, September 2014)

A Technique of Human Action Recognition Using 3D Convolutional Neural Network

Anoosha. P[1], Nandakumar. P[2]
[1]M.Tech student, [2]Professor, NSS College of Engineering, Palakkad, India

***Abstract*— This research involves the study of 3D Convolutional Neural Network (CNN) and its application as human action recognition from a video data base. Human action recognition is very important in real world environment. The current methods used for action recognition are highly problem dependent. Therefore the automated method known as 3D convolutional Neural Network is used for human action recognition in the field of image processing. Here a database is created which consists of features, trained, 3D CNN features were extracted using hardware layered architecture by performing 3D convolution and sub sampling separately on each channel .Thus multiple features are extracted. Number of actions limited was three. The main application of this method is in the field of video surveillances, to find the presence or absence of cluttered backgrounds, occlusions etc.**

***Keywords*— 3D Convolution, Architecture of 3D CNN.**

I. Introduction

Recognizing human action is very important and interesting in the field of video surveillance, shopping behaviour analysis etc. Most of the existing techniques are problem dependent and very complicated to implement. Therefore a new approach named as 3D convolutional neural network for human action recognition was proposed. It is a supervised learning approach. Convolutional neural networks are a type of hardware models that can extract features from video input streams directly. Complex features can be extracted using 3D CNN. The performance analysis is performed by using KTH dataset, which consists of 6 biological actions such as clapping, running, boxing, hand waving, walking and jumping. In which the input video is applied to the CNN layers directly. A kernel with a defined size is used to convolve the input data stream to get multiple channels of information. The multiple channels are gray, gradient x, gradient y, optical flow x, optical flow y. Then alternate sub sampling and convolution are performed to get different numbers of feature Maps in each location in multiple channels. Finally all the outputs from these multiple channels are combined to get final feature representation. Back propagation is the learning algorithm used to train the input video in neural network. This model has been modified to recognize actions in video data.

Deep learning models are a class of machines that can learn a hierarchy of features by building high level features from low-level ones, thereby automating the process of feature construction. Such learning machines can be trained using either supervised or unsupervised approaches, and the resulting systems have been shown to yield competitive performance in visual object recognition, natural language processing and audio classification tasks.

Fig.1.pedestrian actions

II. Convolutional Neural Network

Convolution is a mathematical operation on two functions f and g, producing a third function that is typically represented as a modified version of one of the original functions, giving the area overlap between the two functions as a function of the value that one of the original functions is translated. The convolution of f and g is written as f*g, using an asterisk or star. It is defined as the integral of the product of the two functions after one is reversed and shifted. As such, it is a particular kind of integral transform.

Lecun et al[1] proposed an advanced convolutional architecture.

Neural network refers to the circuit or network of human neurons. It has a remarkable ability to extract patterns. It processes similar way that of human brain does. The main drawback of its operation is unpredictable.

Main applications are in the field of medicine. Neural network made up of different layers. These layers are input layer, hidden layer and output layer. The inputs are taken by the input layer which is the input of hidden layer. Hidden layer provides further processing on input data. Output layer generates the output corresponding to the input. The main principle used by the Neural network is Back propagation.

A. 2D Convolution

Another name for 2D convolution is "shared weight neural networks". It is done by using Kernels, an array of weights. Sharing of weights-decrease the number of variables and hence it increases the overall performance of the system. Feature maps (multiple planes) are usually used in each layer are introduced so that multiple features can be detected. The main principle used is back propagation. The main drawback of 2D convolution is that it is the most time consuming part of an application. It can extract spatial features from an image.

In 2D convolution weight sharing takes place. Weight sharing done by applying a convolution matrix or Kernel, which is distributed entire location of input data on the input data. The multiple planes used are known as Feature Maps. From this feature map, multiple features are extracted. It is applied only in the case of image dataset to capture spatial features.

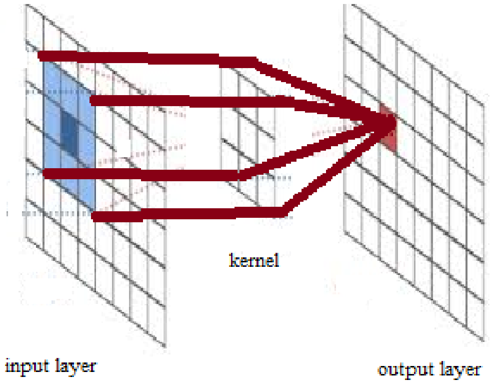


Fig.2.2D convolution

B. 3D Convolution

2D convolution can compute the features from spatial dimension only. So it is enough for still images.

But in the case of 3D or video processing motion information or moving features has to be captured. So computation of both spatial and temporal dimensions is required. To perform this 3D convolution is introduced.

To perform this 2D convolution feature maps to compute features from the spatial dimensions only. When applied to video analysis problems, it is desirable to capture 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 map in the convolution layer is connected to multiple contiguous frames in the previous layer.

C. 3D convolution architecture

To perform this 2D convolution feature maps to compute features from the spatial dimensions only. When applied to video analysis problems, it is desirable to capture 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 map in the convolution layer is connected to multiple contiguous frames in the previous layer.

This architecture is described by Shuiwang ji[2].In this architecture 9 frames of input data with size 60×40 pixel given to the convolution layer. In the first layer a kernel with size 7×7×3 is applied to convolve the input data. The convolved output creates multiple channels of information. It results 33 Feature Maps in 5 different channels. Five different channels are gray, gradient x, gradient y, optflow x and optflow y. Gray channel represents image with gray pixels. Gradient is the directional change in the intensity or colour in an image. Gradient x represents intensity change in x direction and gradient y represents intensity change in y direction .Optical flow provides the motion between two image frames in either horizontal or vertical directions. This hardwired layer computes the features.

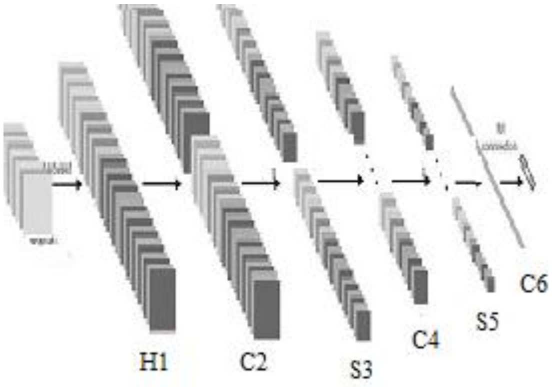


Fig.3.3D CNN architecture

Since people detection is a important role in action recognition, we added an efficient people detector. The people detector object detects people in an input image using the Histogram of Oriented Gradient (HOG) features [3] and a trained Support Vector Machine (SVM) classifier [4]. The main advantage of HOG/SIFT representation is that it captures edge or gradient structure which is very characteristic of local shape, and it does so in a local representation with an easily controllable degree of invariance to local geometric and photometric transformations, translations or rotations make little difference if they are much smaller that the local spatial or orientation bin size. We have shown that using locally normalized histogram of gradient orientations features similar to SIFT descriptors [5] in a dense overlapping grid gives very good results for person detection, reducing false positive rates by more than an order of magnitude relative to the best Haar wavelet based detector[6].

The 3D CNN algorithm deals with gray images (monochrome). So as to improve the performance we first covert rgb image into ycbcr colour space and only luminance component is selected for CNN cause other two components represent colour.

Also execution speed of testing stage is an another important factor , which depend on the size of feature vectors. So to reduce size of feature vectors or extract principle components of feature vectors we use PCA algorithm [7]. Histogram projection method of feature extraction used for extracting feature pixels from the input image and then afterwards how to use Principal Component Analysis (PCA) to reduce the size of feature vector.

III. Simulation Results

A. Action Recognition

The entire document should be in Times New Roman or Times font. Type 3 fonts must not be used. Other font types may be used if needed for special purposes.

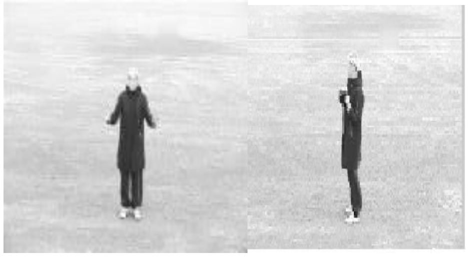


Fig.4.hand clapping and boxing operations

The resultant obtained is

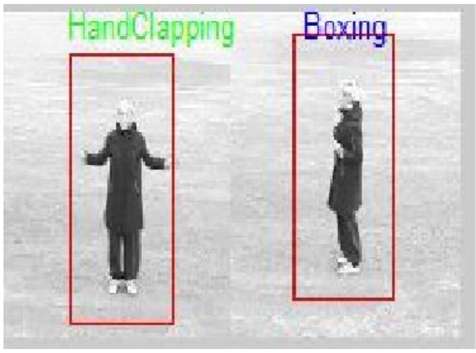


Fig.5.convolved output

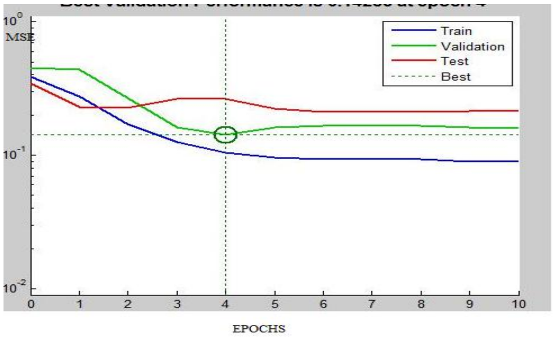
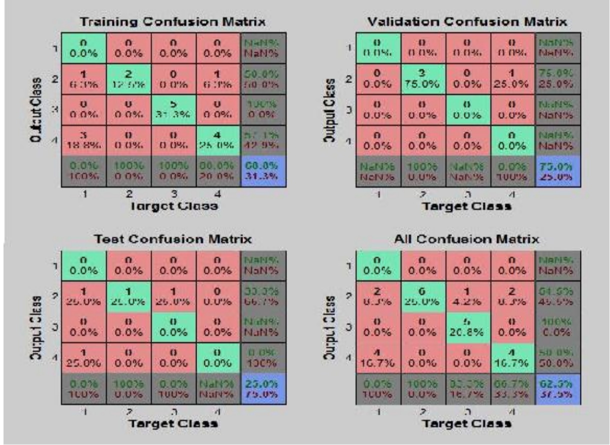


Fig.6.performance curve

The performance curve obtained is given above.



Training Confusion Matrix

	1	2	3	4	
1	0 0.0%	0 0.0%	0 0.0%	0 0.0%	NaN% NaN%
2	1 6.3%	2 12.5%	0 0.0%	1 6.3%	50.0% 50.0%
3	0 0.0%	0 0.0%	5 31.3%	0 0.0%	100% 0.0%
4	3 18.8%	0 0.0%	0 0.0%	4 25.0%	57.1% 42.9%
	0.0% 100%	100% 0.0%	100% 0.0%	80.0% 20.0%	68.8% 31.3%

Output Class
Target Class

Validation Confusion Matrix

	1	2	3	4	
1	0 0.0%	0 0.0%	0 0.0%	0 0.0%	NaN% NaN%
2	0 0.0%	3 75.0%	0 0.0%	1 25.0%	75.0% 25.0%
3	0 0.0%	0 0.0%	0 0.0%	0 0.0%	NaN% NaN%
4	0 0.0%	0 0.0%	0 0.0%	0 0.0%	NaN% NaN%
	NaN% NaN%	100% 0.0%	NaN% NaN%	0.0% 100%	75.0% 25.0%

Output Class
Target Class

Test Confusion Matrix

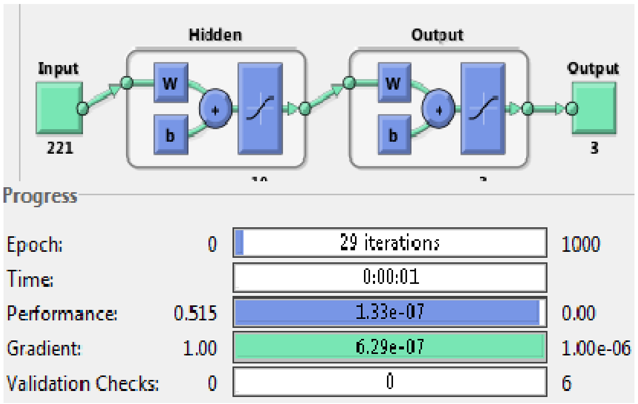
	1	2	3	4	
1	0 0.0%	0 0.0%	0 0.0%	0 0.0%	NaN% NaN%
2	1 25.0%	1 25.0%	1 25.0%	0 0.0%	33.3% 66.7%
3	0 0.0%	0 0.0%	0 0.0%	0 0.0%	NaN% NaN%
4	1 25.0%	0 0.0%	0 0.0%	0 0.0%	0.0% 100%
	0.0% 100%	100% 0.0%	0.0% 100%	NaN% NaN%	25.0% 75.0%

All Confusion Matrix

	1	2	3	4	
1	0 0.0%	0 0.0%	0 0.0%	0 0.0%	NaN% NaN%
2	2 8.3%	6 25.0%	1 4.2%	2 8.3%	54.5% 45.5%
3	0 0.0%	0 0.0%	5 20.8%	0 0.0%	100% 0.0%
4	4 16.7%	0 0.0%	0 0.0%	4 16.7%	50.0% 50.0%
	0.0% 100%	100% 0.0%	83.3% 16.7%	66.7% 33.3%	62.5% 37.5%

Fig.7. confusion matrix obtained

The confusion matrix determines the success ratio of the experiment.



Progress

Epoch:	0	29 iterations	1000
Time:		0:00:01	
Performance:	0.515	1.33e-07	0.00
Gradient:	1.00	6.29e-07	1.00e-06
Validation Checks:	0	0	6

Fig.8.neural network training output

REFERENCES

[1] LeCun .Y.,Bottou.L.,Bengio.Y. and Hafner.P.1998.Gradient based learning applied to document recognition.Proceedings of the IEEE,86(11):22782324.

[2] Shuiwang Ji, WeiXu Ming Yang and Kai Yu. jan.2013"3D Convolutional NeuralNetworks for human action Recognition" IEEE Trans.Pattern Analysis and Machine Intelligence vol.35 no.2

[3] Navneet Dalal.Bill Triggs,Histograms of oreanted gradients for human detection, INRA

[4] T. Joachims. Making large-scale svm learning practical. In B. Schlkopf, C. Burges, and A. Smola, editors, Advances in Kernel Methods - Support Vector Learning. The MIT Press, Cambridge, MA, USA, 1999

[5] Lowe. Distinctive image features from scale-invariant keypoints,I JCV, 60(2):91.110,2004

[6] .Krishnakant C. Mule Anilkumar N. Holambe, hand gesture recognition using PCA andhistogram proection,I nternational Journal on Advanced Computer Theory and Engineering (IJACTE).

[7] Y. Murali Mohan Babu,Dr. M.V. Subramanyam,Dr. M.N. Giri Prasad,"PCA basedimage denoising",signal processing,An International Journal (SIPIJ) Vol.3, No.2, April 2012