January-February 2020
ISSN: 0193-4120 Page No. 6051 - 6055

3d CNNs for Pose Classification using KB Dataset

B. Gnana Priya
Assistant Professor
Department of Computer Science and Engineering
Annamalai University

Dr. M. Arulselvi
Associate Professor
Department of Computer Science and Engineering
Annamalai University

Article Info
Volume 82
Page Number: 6051 - 6055
Publication Issue:
January-February 2020

Article History
Article Received*: 18 May 2019*
Revised*: 14 July 2019*
Accepted*: 22 December 2019*
Publication*: 29 January 2020*

Abstract
Human action prediction and classification is a hot area of research that is employed in various fields. Many of 2D action classification algorithms for images using CNN are developed. But, if we try to apply them on fresh datasets like KB dataset we get an average accuracy prediction due to occlusion and foreshortening of limbs, different orientation and rotation. In this paper, we employed a 3D deep convolutional neural network for human action classification. We capture various stills of the few actions in different views, arrange them together like frames in video and trained a 3D CNN. This gives network the ability to leverage various context of the action that lead to an improved performance and accurate classification of poses by the network.

Keywords; 3D CNN, Classification, Human pose estimation, KB dataset, deep learning.

1. INTRODUCTION

CNN are compact networks that employs convolutional operations that makes use of learnable filters as well as nonlinear activation functions for various classification problems in a hierarchical structures. The network makes a compact representation of the inputs provided and then separates them into different classes for classification based on the objective function. CNN extract complex and abstract features from different regions of the input by stacking and down sampling them. The basic layers used here are: Convolutional Layer**,** Pooling Layer and Fully-Connected Layer. The basic building block of the network is the convolutional layer which employs kernels to find complete features in the image. Convolution operation is an element-wise product followed by sum of two matrices that is employed by the Kernel. Normally, to reduce the parameter size and overall computation pooling layers are inserted between convolutional layers in the network. It also prevents overfitting of network by resizing the input.

2D CNNs considers only a single part of a image as input and the adjacent image matrices for are not considered for processing. But, 3D CNNs uses the voxel information of adjacent matrices that leads to a better feature map prediction and performance. This is achieved by using 3D convolutional kernels that comes with additional cost and increase in parameter numbers. 3D convolutions applies a three dimensional filter to the dataset to calculate features of low level for representations. Their output comes in a three dimensional volume space such as cube or cuboids.

2. RELATED WORKS

3D human poses are estimated from monocular images [1] from 2D images is basically a difficult process than 2D pose estimation. This is due to the huge 3D pose space and the ambiguities due to the non reversible perspective projection. We can generate 3D images using depth maps and can use them for estimating 3D human pose [2] that is proven to be effective. The methods used to estimate 3D human poses using a single image are classified into two types: the first one that estimates 2D poses and use them to find 3D poses and second being the approach that directly learn from features of the image and estimates the 3D poses. [3] and [4] uses the 2D joints available to predict the 3D poses from 2D poses. Chen [5] employs a nearest neighbour search on the 3D pose library which they have build using 2D projections that are enormous to estimate 3D pose from 2D projections. Moreno[6] uses a distance matrix

Published by: The Mattingley Publishing Co., Inc.

January-February 2020
ISSN: 0193-4120 Page No. 6051 - 6055

regression to estimate 3D poses from 2D poses. Xiaohan [7] uses LSTM to find the depth of human joints using their locations in 2D view. Martinez [8] builds a deep network to map 2D joints to 3D space that is simple and efficient. These methods are based on 2D pose estimator and their performance are limited. Few of the approaches that directly estimates the 3D poses from features extracted from images are given in [9], [10], [11], [12], [13].

Recently, there has been vast improvement in this area using CNNs architecture that employs end-to-end mappings. Gernot[14] proposed Octnet a new representation for 3D data for deep learning. Here, they use unbalanced octrees to partition the space and features are stored to achieve high and deep resolution network. Memory allocation, resolution, deep networks are the main points here. Rana[15] suggested MeshCNN network that employs triangular meshes. MeshCNN employs special layers that fuses convolution and pooling to work on the edges. Convolutions operate on edges and pooling collapses edges to retain surface topology. The network expands the needed features and discards the features that are redundant. Swalpa [16] uses Hybrid Spectral Convolutional Neural Network for image classification. They employ a 3D-CNN along with 2D-CNN to learn abstract level spatial representations. This hybrid CNN has a reduced complexity for HSI classification and is applied to Indian Pines, Salinas Scene remote sensing and Pavia University datasets.

Huang[25] introduces the Dense layers in CNN that connects every layer to every other layer in a feed forward method. Advantages of DenseNet are they decrease parameters, reuse of features, more propagation of features and alleviate the vanishing-gradient problem. Yongheng[19] proposes a 3D point-capsule networks. This network is developed using an auto-encoder to work on sparse 3D point clouds which preserves the spatial location of the input data. They have achieved a better performance for 3D point cloud related task such as object classification, reconstruction and segmentation of parts through the dynamic routing and 2D latent space deployment. Lars [20] presented a Occupancy Network, that is based on 3D reconstruction methods. This network represents 3D surface as continuous decision boundary of network classifier and provides 3D output with high resolution and minimal use of memory. The network can be applied for 3D pose construction from single image, coarse discrete voxel grids and noisy point clouds. The occupancy networks used to represent high resolution meshes.

Shikun [23] proposes VSL(Variational Shape Learner),to learn 3D shapes in an unsupervised manner. Skip connections are employed to learn hierarchical representation of objects. VSL latent probabilistic manifold method used to generate 3D objects from 2D images. Variational Shape Learner is build based on neural statistician and the convolutional neural network. Kang[24] presented a dual path network that fuses the advantages of dense and residual network. It takes into account the feature redundancy which is not considered by the previous models. In order to extract the distinguishing features from complex scene in HSI, a dual path network (DPN) is proposed for HSI classification. Some other existent approaches that handle 2D and 3D pose estimation together or iteratively [26], [27], [28], [29]. Tekin [28] the network learns how to combine 2D and 3D image cues. Tome [29] uses a CNN that has multi stage and uses the knowledge of possible 3D prompt locations to improve the search for a better 2D locations. CNN architectures for human pose estimation requires a mammoth training data to train the network and to achieve state-of the-art performance. For 2D pose estimation data can be obtained by physically annotating images captured whereas for 3D poses we need to use motion capture (MoCap) systems in restricted environment.

3. PROPOSED METHOD

The proposed method is applied on the KB dataset which contains various poses from Karate and Bharathanatyam. We have collected fresh images and augmented them to get around 3500 images which forms the KB dataset. It has 20 poses, 10 from Karate and 10 from Bharathanatyam. Only single subject is present in each image. The images are taken such that our network can recognize foreshortening and occlusion of limbs, orientation and rotation of the same image. Various views of the same pose are collected in this dataset. This problem is a multi-class classification problem. Previously we trained a 2D CNN that takes a 200 X 200 greyscale image as input. The output is a vector of numbers that indicate the probabilities of each of the 20 poses and we achieved a overall accuracy of 65% only. Since, CNNs need a large set of training data it becomes difficult for us to get a overall increase in accuracy. Here, different views (eight in each) of the same poses are arranged in random order that forms a single patch and are feed to 3D CNN. Our 3D CNN takes training data with dimensions 256 X 256 X 8 as input. The network is build upon our already trained 2D CNN which provides us with great advantage.

We built the network using Keras API which has features that allow us to extend our network or add additional modules effortlessly. A Sequential model where one layer can be added above the other is employed. We use Conv3d and maxpool3d layers to build our network. 3D

Published by: The Mattingley Publishing Co., Inc.

January-February 2020
ISSN: 0193-4120 Page No. 6051 - 6055

convolutions employs a 3D filter to the single patch of images that moves in three directions for abstraction of features in the image. Since the depth is only 8, a stride of one is used for depth dimension in Maxpooling layer. To learn non-linear decision boundaries ReLU activation function is used. Since it is a multiclass classification problem SoftMax is used as the final layer. Fig(1) shows the complete architecture and layers of the network. Dropout layer used to prevent overfitting and ensure that the parameters of the network are not getting biased towards training data. It will drop random connections during our training and dependency of training set may get reduced. Adam optimizer used to configure the network. For multiclass classification the loss type used is categorical cross entropy. The accuracy and loss are the metrics we are tracking during the training process.

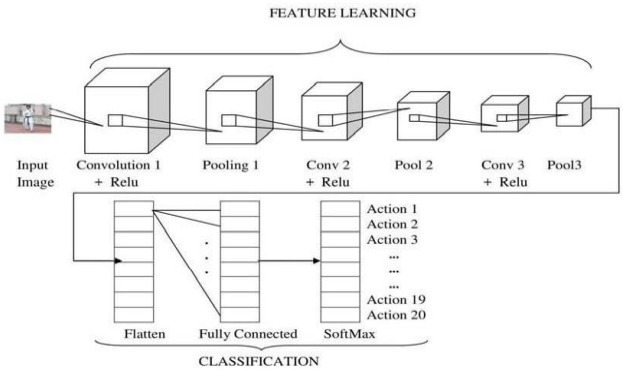
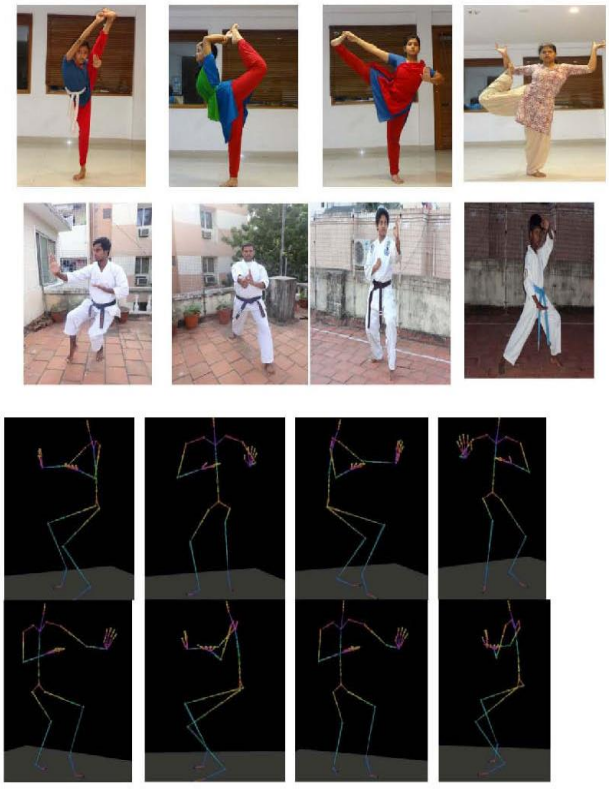


Fig (1) Architecture of the Network



Fig(2) Above- Sample images from KB Dataset: Below-Various view of a single pose

Published by: The Mattingley Publishing Co., Inc.

January-February 2020
ISSN: 0193-4120 Page No. 6051 - 6055

4. RESULTS AND DISCUSSION

The 3D CNNs shows a better performance since they are trained with different views of the same images in a single set. Fig(2) shows sample of different views for a single image. It shows the keypoints ie the joints to describe how the network uses them for classification. It is only imaginary since CNNs do have the ability to learn feature extraction directly. When compared to 2D CNN action classification for the same dataset we have achieved better results. Moreover, 3D CNNs can also be applied for image classification also until we give a three dimensional input.

5. CONCLUSION

In this work we have collected few Karate and Bharathanatyam poses and performed 3D action classification. The classification is carried out based on deep learning algorithm using Keras library running in top of Tensorflow. We use our already pre-trained 2D model and developed a 3D CNN over it for action classification. In our previously trained 2D model classification of certain poses were poor and achieved a overall accuracy of 62%. The 3D CNN gives a accuracy of 85% and some of the poses are well recognized by the network irrespective of the view and occlusion. In future, we plan to build a multi-view dataset including actions from martial, sports and dance for various poses. In future the work can be extended for classifying many diverse poses.

6. REFERENCES

[1] Agarwal, Triggs, " Recovering 3d human pose from monocular images", IEEE Trans. Pattern Analysis, 2006.

[2] Shotton,, Fitzgibbon, Cook, Sharp, " A Real- time human pose recognition in parts from single depth images", CVPR, 2011.

[3] Akhter and Black, “Pose-conditioned joint angle limits for 3D human pose reconstruction”, CVPR, 2015.

[4] Fan, Zheng, Zhou, and Wang, “Pose locality constrained representation for 3D human pose reconstruction” , ECCV, 2014.

[5]Chen and Ramanan, “3D human pose estimation = 2D pose estimation + matching” , CVPR, 2017.

[6] Moreno Noguer, “ 3D human pose estimation from a single image via distance matrix regression” , CVPR, 2017.

[7] Xiaohan Nie, Wei, and Zhu, “Monocular 3D human pose estimation by predicting depth on joints”, ICCV, 2017.

[8] Martinez, Hossain,. Romero, and Little, “A simple yet effective baseline for 3D human pose estimation”, ICCV, 2017.

[9] Agarwal.A and Triggs, “3D human pose from silhouettes by relevance vector regression”, CVPR, 2004.

[10] Rogez, Rihan, Ramalingam, “Randomized trees for human pose detection”, CVPR, 2008.

[11] Sminchisescu, Kanaujia, Li, and Metaxas, “Generative modeling for continuous non-linearly embedded visual inference” , CVPR, 2005.

[12] Bo, Sminchisescu, Kanaujia, and Metaxas, “Fast algorithms for large scale conditional 3D prediction”, CVPR, 2008.

[13] Shakhnarovich, Viola, and Darrell, “Fast pose estimation with parameter-sensitive hashing”, CVPR, 2003.

[14] Gernot , Ali and Andreas , “OctNet: Learning Deep 3D Representations at High Resolutions”, Computer Vision and Geometry Group.

[15] Rana ,Amir, Fish, “MeshCNN: A Network with an Edge”, arXiv:1809.05910v2, 13 Feb 2019.

[16] Swalpa , GopalKrishna, ShivRam , and Bidyut , “HybridSN: Exploring 3D-2D CNN Feature Hierarchy for Hyperspectral Image Classification”, IEEE (Doi: 10.1109/Lgrs.2019.2918719) .

[17] Daniel, Ronald and Hao , “Extending Adversarial Attacks and Defenses to Deep 3D Point Cloud Classifiers”, arXiv:1901.03006v4, 28 Jun 2019.

[18] Lin Shao, Peng Xu, “ Neural Network for 3D object classification”, Stanford University, 94305, CA.

[19] Yongheng , Tolga l, Haowen Deng and Federico , “3D Point-Capsule Networks”, arXiv:1812.10775v1 [cs.CV] 27 Dec 2018.

[20] LarsMescheder, Michael, Michael.N , Sebastian N and Andreasr, “Occupancy Networks: Learning 3D Reconstruction in Function Space” Autonomous Vision Group, MPI for Intelligent Systems, April 2019.

[21] Hicham , Mostafai, Amal ,and Ahmed “ Classification and Recognition of 3D Image of Charlier moments using a Multilayer Perceptron Architecture”, Proceedings of Computer Science 127, 2018.

[22]Asako , Yasuyuki and Yoshifumi, “RotationNet: Joint Object Categorization and Pose Estimation Using Multiviews from Unsupervised Viewpoints”, arXiv:1603.06208v4 , 2018.

[23] Shikun , Lee Giles and Alexander , “Learning a Hierarchical Latent- Variable Model of 3D Shapes”, arXiv:1705.05994v4 , 2018.

[24] X. Kang, Zhuo, and Duan, “Dual-path network-based hyperspectral image classification,” in IEEE Geoscience and Remote Sensing Letters, 2018.

Published by: The Mattingley Publishing Co., Inc.

January-February 2020
ISSN: 0193-4120 Page No. 6051 - 6055

[25] Huang, Liu, van der Maaten, and. Weinberger, "Densely connected convolutional networks," in Proceedings of IEEE Conference, 2017.

[26] Simo Serra, Quattoni, Torras, and Moreno-Noguer, "A joint model for 2D and 3D pose estimation from a single image" ,CVPR, 2013.

[27]Zhou and La Torre, "Spatio-temporal matching for human detection in video" , ECCV, 2014.

[28] Tekin, Marquez, Salzmann, and Fua, "Learning to fuse 2D and 3D image cues for monocular body pose estimation", ICCV, 2017.

[29] Tome, Russell, and Agapito, "Lifting from the deep: Convolutional 3D pose estimation from a single image", CVPR, 2017.

[30]Mehta, Rhodin, Casas, "Monocular 3D human pose estimation in the wild using improved cnn supervision" ,2017.

Published by: The Mattingley Publishing Co., Inc.