Deep Convolutional Neural Network for Human Pose Estimation

B. Gnana Priya

Assistant Professor, Department of Computer Science and Engineering, Annamalai University

Dr. M. Arulselvi

Assistant Professor, Department of Computer Science and Engineering, Annamalai University

ABSTRACT

Human pose estimation is the basic key element which makes us to understand what people do in an image or a given video. From the pose estimated we can recognise the action or action sequences. This is achieved by estimating the key points in the body and thereby predicting the posture of the person. Instead of using the benchmark datasets available here our own dataset which contains karate action poses are used. Deep convolutional neural networks that are more popular nowadays used for pose estimation.

KEYWORDS: Human pose estimation, Deep learning, Convolutional neural network, Karate dataset

1. INTRODUCTION

Human pose estimation is one of the most challenging and popular area of research. The various applications include surveillance, autonomous driving vehicle, gaming, human computer interaction, assisted living, activity recognition, image indexing and retrieval and so on. Human pose plays an important role in recognizing the actions and predicting them. Action recognition is usually performed once the action is complete. In certain cases we need to predict the entire action using few poses performed initially in a video and is known as action prediction. Human action classification is needed in order to recognize the actions that humans are performing. Understanding human behaviour and their interaction with surroundings are very much essential and are in demand for industrial applications.

Karate is an unarmed martial-arts discipline which contains defensive and counter attacking body movements. Karate employs kicking, striking, and defensive blocking with arms and legs. Striking surfaces include the hands, ball of the foot, heel, forearm, knee, and elbow. KATA is a formalized sequence of movements which represent various offensive and defensive postures. These poses are captured and are used to train our classifier.

1.1 Convolutional Neural Network

CNN is a hierarchical learning which allows us to automatically learn features from training process instead of hand designed feature extraction process. Hand designing a set of rules and algorithms to extract features from an image were earlier employed and is a tough and time consuming process. The pixel intensity values of an image serves as input to CNN. A series of hidden layers are used to extract features of the given image. The lower level layers to higher levels are used to extract simple to complex or more abstract features. In lower

level layers simple features like edges are detected. The intermediate layers combine the simple features found previously and finds the corners and outline of the objects. Higher level layers (layers at the end) combine the edges, corners and outlines to form abstract objects.

Three types of layers are used here viz Convolutional Layer**,** Pooling Layer and Fully-Connected Layer. Convolutional layer forms the basic building block and uses kernels to detect features all over the image. The Kernels carries out a convolution operation which is an element-wise product and sum between two matrices. Pooling layers are inserted between convolutional layers to reduce the parameters and computation in the network thereby speeding up the training process on large data. It also resizes the input and prevent overfitting of network. Weight sharing method is used to speed up the training process on new set of data and improves performance of CNN.

2. RELATED WORKS

Earlier human pose was estimated using sliding window part detectors from a set of learned feature [2,3]. Later parameter sensitive hash function used for learning and perform example based pose estimation[4].This were performed on controlled recording environments with defined conditions. Then set of features were extracted in unconstrained environments using regression based methods, nearest neighbour-hood or support vector machine based architectures. Edge based histograms [5] and silhouette features [6] are also employed. Pictorial structures [7, 8], poselets [9] and part models [10] are also used.

Human pose estimation using CNN's can be classified as detection based and regression based methods. Deep Convolutional Neural Network(DCNN) have achieved a significant improvement in human pose estimation[1,11,12,13,19,22,24]. Most DCNN follow the method of regressing heatmaps of each body parts. The regression model has the ability of learning feature representations. In order to improve performance regression methods applied in a sequential, cascaded fashion have been mostly used now.[26] have achieved outstanding results in both LSP and MPII datasets using a six stage CNN cascade.

It will be very difficult for DCNN's to regress accurate heatmaps for body parts with heavy occlusions and cluttered background. The point here is that we need a large set of training data to learn the real body joints distribution. Instead of learning explicitly some models attempt to learn human body structures implicitly. GAN's [27, 28] which are recently very popular approach has a generator (generates new instances) and the discriminator (evaluates for authenticity).

3. PROPOSED WORK

We use deep convolutional neural network to perform the classification of the actions performed by humans. Instead of taking the normal day to day actions we use actions taken from karate a form of martial arts where the poses are different and rare.

The data was recorded in a karate school in a open atmosphere with cluttered background. The data was captured using three cameras placed around the capture space. Twenty different persons performed the same actions and are captured. The poses are from different parts of the action sequence while doing karate Kata. We have chosen 500 images from the original captured image which are clear enough for processing. Each pose is captured in different angle so that our system can recognize foreshortening of limbs, occlusion of limbs, rotation and orientation of the figure. Data augmentation is done to increase the number of images needed for training. The captured images are rotated to different angles from the original angle. We got around 768 images after augmentation. Each of our image contains only single subject and are centred in the image.

Action Category	No. Of Actions	No. Of Images
KARATE	1.Karate Action 1	100
	2.Karate Action 2	100
	3.Karate Action 3	100
	4.Karate Action 4	100
	5.Karate Action 5	92
	6.Karate Action 6	92
	7.Karate Action 7	92
	8.Karate Action 8	92
Total		768

Table 1: Total Number of images in the Karate Dataset

The various steps carried out are

- Capturing the Data
- Pre- processing the Dataset.
- Modelling the Convolutional neural network using Keras.
- Training the CNN for multiclass classification.
- Evaluating the model and predicting the output class of a test image.
- Finding the accuracy and loss.
- Plotting the Confusion Matrix.

4. NETWORK ARCHITECTURE

We use Keras API written in python. Keras specially designed for neural network running on top of TensorFlow. It allows us to built networks easily, extend them and add new modules in a simple manner.

We use the Sequential model for building our network. The desired layer can be added one by one in the Sequential model. The Dense layer (fully connected layer) is used to build a feed forward network in which all the neurons from one layer are connected to the neurons in the previous layer. ReLU activation function is required to give non-linearity to the model. This will make the network to learn non linear decision boundaries. As our problem is a multiclass classification problem we use the SoftMax layer as the final layer.

In order to configure the network we use SGD (Stochastic Gradient Descent) optimizer. The loss type used here is categorical cross entropy which is used for multiclass classification. The accuracy and loss are the metrics we are tracking during the training process. The Confusion matrix for the various action categories are plot. We can infer the percentage of correct recognition of a particular pose compared with other poses through this matrix.

Fig 1: Sample images from karate Dataset

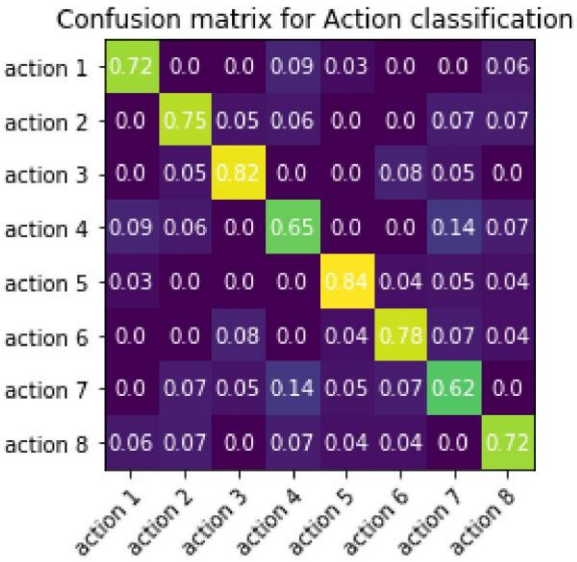

Fig 2: Confusion Matrix for Action Classification

5. CONCLUSION

Today Convolutional neural networks are the driving force behind machine learning and computer vision applications like robotics, medical diagnostics, self driving cars, etc.. This is because of their ability to work with few parameters and their simplicity. Deep learning algorithm for action classification for few karate moves using Keras library running in top of Tensorflow is employed. The proposed work classifies the pose with an accuracy of 73% .In future this work will be extended for videos. This image classifier will serve as basic entity for video classification of various karate movement sequences. Also, the network accuracy can be increased and the network can be trained for classifying complex and occluded diverse poses.

6. REFERENCES

1) Li Wei and Shishir K. Shah , Human Activity Recognition using Deep Neural Network with Contextual Information In Proceedings of the 12th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2017), pages 34-43
2) V.Athitsos, J.Alon, S.Sclaroff and G.Kollios, Bootmap: A method for efficient approximate similarity rankings, CVPR, 2004
3) A.Farhadi, D.Forsyth and R.White, Transfer learning in sign language, CVPR,2007.
4) G.Shakhnarovich, P.Viola and T.Darrell , Fast pose estimation with parameter sensitive hashing, ICCV, 2003.
5) G.Mori and J.Malik, Estimating human body configurations using shape context matching, ECCV,2002.
6) K.Grauman, G.Shakhnarovich and T.Darrell, Inferring 3d structure with a statistical image based shape model, ICCV , 2003
7) B.Sapp, C.Jordan and B.Taskar, Adaptive pose priors for pictorical structures, CVPR, 2010
8) L.Pishchulin, A.Jain, M.Anduriluka, T.Thormaehlen and B.Schiele, Articulated people detection and pose estimation, Reshaping the future, CVPR, 2012
9) L.Bourdev and J.Malik, Poselets: Body part detectors trained using 3d human pose annotations, ICCV, 2009
10) P.Felzenszwalb, D.Ramanan, A Discriminatively trained, multiscale, deformable part model, CVPR, 2008
11)Mona M. Moussa , Elsayed Hamayed b, Magda B. Fayek b, Heba A. El Nemr , An enhanced method for human action recognition, Journal of Advanced Research (2015) 6, 163–169
12) Weichen Zhang, Zhiguang Liu, Liuyang Leung, Howard Leung, Martial arts, Dancing and Sports Dataset: A challenging Stereo and multiview dataset for 3D Human pose estimation, Image and Vision computing, Feb 2017
13) Yu Kong, Member, IEEE, and Yun Fu, Senior Member, Human Action Recognition and Prediction: A Survey ,IEEE journal of latex class files, vol. 13, no. 9, september 2018
14) J. Donahue, L. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in CVPR, 2015.
15) A. Kar, N. Rai, K. Sikka, and G. Sharma, "Adascan: Adaptive scan pooling in deep convolutional neural networks for human action recognition in videos," in CVPR, 2017.
16) Y. Yang and M. Shah, "Complex events detection using data-driven concepts," in ECCV, 2012.
17) K. Wang, X. Wang, L. Lin, M. Wang, and W. Zuo, "3d human activity recognition with reconfigurable convolutional neural networks," in ACM Multimedia, 2014.
18) G. W. Taylor, R. Fergus, Y. LeCun, and C. Bregler, "Convolutional learning of spatio-temporal features," in ECCV, 2010.
19) L. Sun, K. Jia, T.-H. Chan, Y. Fang, G. Wang, and S. Yan, "Dl-sfa: Deeply-learned slow feature analysis for action recognition," in CVPR-2014
20) T. Pl otz, N. Y. Hammerla, and P. Olivier, "Feature learning for activity recognition in ubiquitous computing," in IJCAI, 2011.
21) Q. V. Le, W. Y. Zou, S. Y. Yeung, and A. Y. Ng, "Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis," in CVPR, 2011.
22) S. Ji, W. Xu, M. Yang, and K. Yu, "3d convolutional neural networks for human action recognition," in ICML, 2010.
23) M. Hasan and A. K. Roy-Chowdhury, "Continuous learning of human activity models using deep nets," in ECCV, 2014.
24) Alexander Toshev, Christian Szegedy, DeepPose: Human Pose Estimation via Deep Neural Networks .
25) Alex Krizhevsky, University of Toronto, Ilya Sutskever, Geoffrey E. Hinton ,ImageNet Classification with Deep Convolutional Neural Networks.
26) Wei, S.E.Ramakrishna, V.Kanade, T.Sheikh, Convolutional pose machines, CVPR, 2016
27) A.Raford, L.Metz and S.Chintala, Unsupervised representation learning with deep convolutional generative adversarial networks, arXiv, 2015
28) T.Salimans, I.J.goodfellow, W.Zaremba, V.Cheung, Improved techniques for training GAN's, In Proc. Advances in neural Inf. process systems ,2016