Application of Convolutional Neural Network to Classify Sitting and Standing Postures

Nishank Singhal, Srishti, and V. Kalaichelvi

***Abstract*—The paper aims at identifying whether a sitting or standing posture of a person is correct or incorrect using image processing and deep learning approach. The approach includes: (i) to check whether a person is present or not in the image read (ii) if present, then detect whether the person is sitting or standing (iii) in case of sitting, identify whether the posture is correct or incorrect (iv) in case of standing, identify whether the posture is correct or incorrect. In accomplishing the task, an overall accuracy of 91.3% is achieved. The method has been evaluated by testing it with a real time video feed thereby demonstrating the efficiency of the model and the wonderful power of Convolutional Neural Network (CNN).**

***Index Terms*—image processing, convolutional neural network, posture detection.**

I. Introduction

THE goal of this paper is to estimate a 2D human upper body posture by involving the identifications of the key points of the neck joints as shown in Fig.1. It is important to clarify here what the term posture means as different professionals define it differently. According to Yasmeen, Shahrukh, and Farooqui, posture is the position in which one holds their body against gravity while sitting or standing [1].

With change in times, where spine and neck related problems were mainly associated with adults a decade ago, today are also associated with school-aged and university-aged students. One of the causes is having poor posture [2] often about which students are not aware [3]. Even young computer professionals due to their sedentary lifestyle fall victim to the pain which may be trivial in the early stages of their life, but may grow as they grow older because of their bad posture [4]. Occupations like the above, which involve overusing the neck muscles and staying in one posture (which by habit is usually incorrect) continuously for long periods of time may lead to the condition of cervical spondylosis. Also known as the arthritis of the neck, cervical spondylosis is a condition of the spine which leads to bouts of stiffness and neck pain [5].

The idea here is to classify one's posture as it has a major impact on one's life. In carrying out the whole work, Image Processing and Convolutional Neural Networks (CNN), both play an equally integral role.

With its forward-looking and refined ways, image processing has achieved remarkable success when it comes to image classification and object localization. In this paper, Image Processing acts as a pre-processing technique responsible for highlighting the region of interest, in this case, the human being in order to provide only the required features to the CNN.

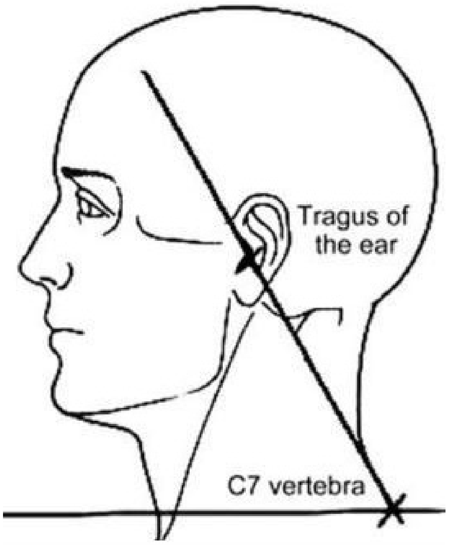


Fig. 1. Side Pose of a Person to Find the Craniovertebral Angle.

As for CNN, the last couple of years have seen a drastic change in the field of computer vision with its futuristic, fast and scalable end-to-end learning framework [6]. CNNs have always been known for their excellent application in image related tasks and are very useful where convolutional constraints are a potential premise. Like any other standard multi layered network, CNNs too comprise of one or more hidden layers along with subsampling steps and fully connected layers, but is able to put itself at a higher spot due to its ease of training with fewer parameters without compromising on the number of hidden layers involved [7].

A. Problem Statement

The two major tasks of the problem involve: estimation of craniovertebral angle and posture classification. Image processing techniques help in identifying the points of interest for finding out craniovertebral angle as shown in the Fig.2. This is followed by employing CNN as a training technique as explained in section 3.

II. Background

A. Image Processing

To highlight the subject, in this case, the human being from a video feed, background subtraction for foreground detection has been used. The method extracts the change in the second image of the two consecutive images in the feed,

Manuscript received July 23, 2017.

Nishank Singhal is a student of BITS, Pilani in the Department of Computer Science, Dubai Campus, UAE. (Corresponding Author: e-mail: nishanksinghal20nov@gmail.com).

Srishti is a student of BITS, Pilani in the Department of Computer Science, Dubai Campus, UAE (e-mail: srishti258@gmail.com).

V. Kalaichelvi is an Asst. Prof. in the Department of Electrical and Electronics Engg., Dubai Campus, UAE at BITS, Pilani (e-mail: kalaichelvi@dubai.bits-pilani.ac.in).

until the end. The individual results so obtained are termed as foreground mask [8], which is used in this work. The new background model, given by $B_{t+1(x,y)}$ for each frame is estimated as:

$$B_{t+1(x,y)} = \alpha I_{t(x,y)} + (1-\alpha) B_{(t+1(x,y))} \quad (1)$$

Here, $I_{t(x,y)}$ is the current pixel value where t is the frame number and (x,y) is the pixel location in the frame and α is the speed of updating background model.

The difference between current frame and background is calculated as:

$$D_{t(x,y)} = |I_{t(x,y)} - B_{t(x,y)}| \quad (2)$$

The pixels whose difference value is greater than given threshold T are classified as foreground pixels given by [9]:

$$|M_{t(x,y)}| = \begin{cases} 0 & \text{if } D_{t(x,y)} \leqslant T \\ 1 & \text{if } D_{t(x,y)} > T \end{cases} \quad (3)$$

Literature offers other methods [10], [11], but this method of foreground detection helps in removing the restriction of color specificity and thereby complementing the methodology followed. Finally, through contouring techniques, points and regions of interests are obtained as shown in Fig.2.

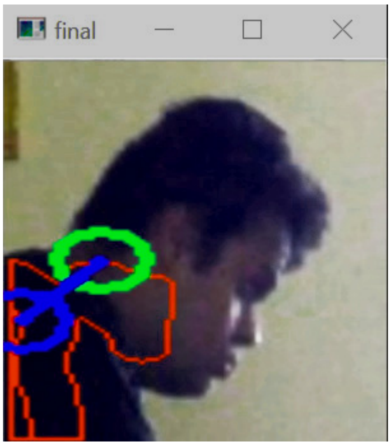


Fig. 2. Identifying the points of interest for finding out craniovertebral angle

B. Convolutional Neural Network

The abstract description of a CNN can be visualized as follows:

$$x^1 \rightarrow w^1 \rightarrow x^2 \rightarrow \cdots \rightarrow w^{L-1} \rightarrow x^L \rightarrow w^L \rightarrow z \quad (4)$$

As (4) suggests, a CNN runs layer by layer in the forward direction. The input, x^1, goes through processing in the first layer, w^1, which is collectively called as a tensor. This layer maybe a convolutional layer, pooling layer, fully connected layer, etc. The process continues until the (L-1)th layer is reached which is usually a softmax layer (to make x^L a probability mass function) followed by a last layer which is the loss layer.

Finally, $x^L \in R^C$ is achieved, i.e. a column vector with C elements, which estimates the posterior probabilities of x^1 belonging to the C categories. The CNN is prediction as [12]:

$$arg_i max(x_i^L) \quad (5)$$

It is worth mentioning here that pooling and subsampling layers help in reducing the noise in an image by decreasing its resolution [13]. Loss layers are responsible for minimizing the cost by computing the loss of the forward pass and the gradient descent w.r.t. the loss of backward propagation. The softmax with loss layer just provides a numerically stable gradient [14]. The work of the convolution operation is to extract features from the input. The convolutional procedure can be expressed mathematically as follows:

$$y_{i^{l+1},j^{l+1},d} = \sum_{i=0}^{H}\sum_{j=0}^{W}\sum_{d^l=0}^{D^l} f_{i,j,d^l,d} \times x^l_{i^{l+1}+i,j^{l+1}+j,d^l} \quad (6)$$

where x is a third order tensor such that $x \in R^{H \times W \times D}$. H, W and D are elements each indexed by an index triplet (i,j,d). $0 \leqslant d \leqslant D = D^{l+1}$, and for any spatial location (i^{l+1}, j^{l+1}) satisfying $0 \leqslant i^{l+1} < H^l - H + 1 = H^{l+1}, 0 \leqslant j^{l+1} < W^l - W + 1 = W^{l+1}$. In this equation, $x^l_{i^{l+1}+i,j^{l+1}+j,d^l}$ refers to the element of x^l indexed by the triplet $(i^{l+1}+i, j^{l+1}+j, d^l)$.

The charm of a CNN lies in the fact that all spatial locations share the same filter which leads to less parameters being involved and subsequently reducing the training time. The sharing of parameters promotes effective learning of good features in images. This means that out of all M features, it is possible that only one feature is useful in recognizing all Y object categories or a joint of all M features is required to recognize an object type of Y. Taking advantage of the same, classifications have been implemented which are discussed in detail in the following section.

C. Cervical Spondylosis

Though the condition is more common among middle aged and elderly people indicating age as a risk factor, occupations that involve long sitting hours and holding tense postures also pose a high risk of the same in the near future. Its symptoms include headache in the back of the head, numbness in arms and shoulders and neck stiffness which may worsen over time.

In this paper, the craniovertebral angle is taken into consideration while classifying a posture as correct or incorrect. The craniovertebral angle is the angle formed at the intersection of a horizontal line through spinous process of C7 and the line of the tragus of the ear as illustrated in Fig.1. This is believed to provide an estimation of neck on upper trunk positioning. A small angle indicates forward head posture [15].

III. Datasets, Implementation and Results

A. Dataset for evaluation

The dataset used in this work spans approximately over 40,000 images in total. They are further divided into four categories, namely person and non-person (background) (about 40,000), sitting and standing (about 22,000), standing: correct and incorrect posture (about 10,000 whose sample is shown in Fig.3) and sitting: correct and incorrect posture (about 10,000 whose sample is shown in Fig.4). As a case study, we consider school-aged and university-aged students and young computer professionals filmed while studying, working on a computer and watching TV. Each image is extracted from a video feed and all are about 384x172 pixels in size.

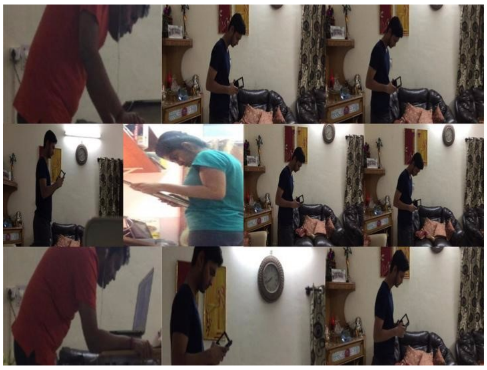


Fig. 3. Sample of the Standing Posture Dataset

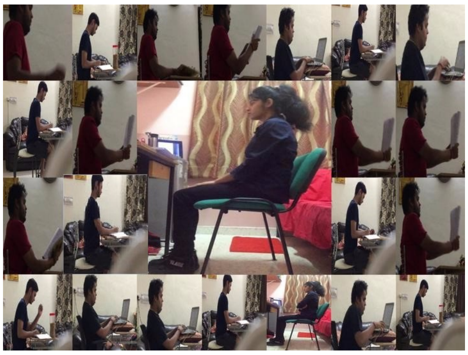


Fig. 4. Sample of the Sitting Posture Dataset

B. Implementation

1) Image Processing: Fig.5 demonstrates the steps of implementation of the whole procedure. To elaborate, video feed is read frame by frame from the dataset. The frame read is then applied for background subtraction using Gaussian Mixture based background/foreground segmentation algorithm using the method of 'BackgroundSubtractorMOG' available in OpenCV. After background subtraction, the foreground object is segmented out. The morphological operation of dilation is then applied in order to remove noise from the segmented frame. After the contouring of the segmented frame is done, the contours of the frame are found out with the help of a function 'findContours'. Out of all the contours found, only the biggest contours are taken into consideration in order to remove the extraneous background contours. The frame so formed is then resized to 400x400 pixels.

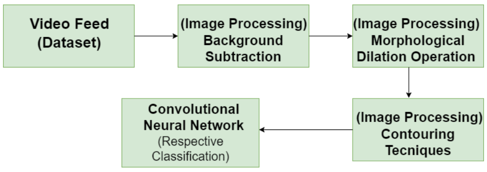

Video Feed (Dataset)
(Image Processing) Background Subtraction
(Image Processing) Morphological Dilation Operation
(Image Processing) Contouring Tecniques
Convolutional Neural Network (Respective Classification)

Fig. 5. Steps of Implementation for Posture Classification

2) Training Detail: With reference to Fig.6, an image of size 80x80 is fed as input into the CNN consisting of four hidden layers and two fully connected layers with a ReLU and softmax as last layer. Each convolutional layer uses a filter of 5x5 and a stride of 2. Each convolutional layer is followed by a max-pooling layer of filter size 5x5 and a stride of 2. The CNN is trained with 20,500 images against a validation set of 1,500. As per the Fig.7, using a batch size of 80, validation is performed over 4.500 iterations. A constant learning rate of 0.001 is used throughout.

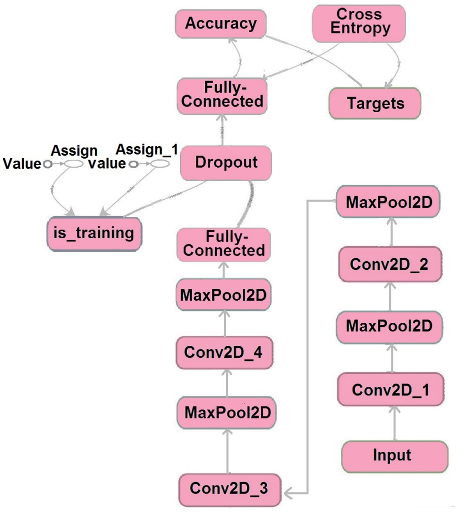


Fig. 6. CNN Architecture for Posture Classification

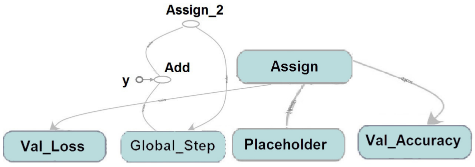


Fig. 7. Structure for Validation of Training Data

3) Testing method: The testing of our classification is done on a validation set of nearly 1,500 images. The testing phase is designed to act as a multi-layered binary classification system. As shown in Fig.8, after obtaining the static images from the video feed and cropping them, the CNN first tests the cropped static images obtained from the video feed for the presence of a person. If present, it then checks whether the person is sitting or standing. Otherwise, it classifies the image as non-person or background. Upon checking, if the person is found to be sitting, the CNN classifies the sitting posture as correct or incorrect posture, otherwise, it classifies the standing posture as correct or incorrect.

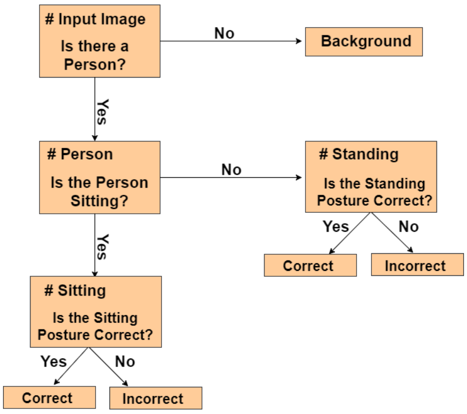


Fig. 8. Procedure for Classification of Posture

C. Results and Discussion

Fig.9 shows the simulation results of accuracy vs the number of training images. It is observed that the testing accuracy is directly proportional to the number of training images. According to the graph in Fig.9, the accuracy increases from 70% to 91%, as the number of training images increase from 4,000 to 22,000. The training images comprise of images depicting sitting and standing postures. The testing images are held constant to 3,000 images.

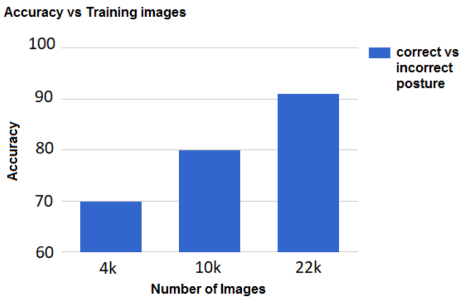
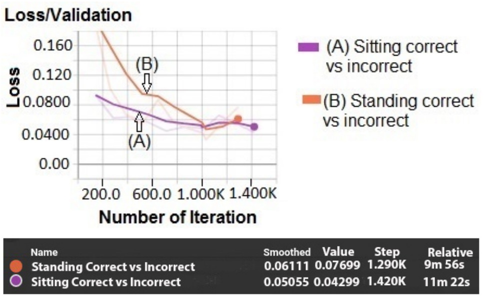


Fig. 9. Testing Accuracy with regard to the Number of Training Images

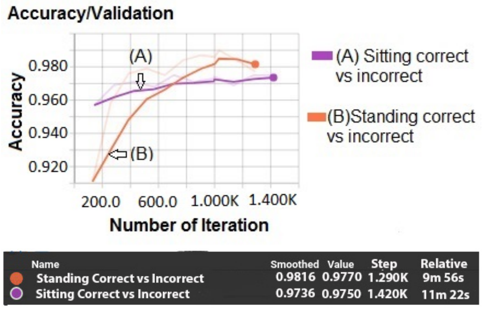
We test our classification CNN on a validation set of approximately 1,500 distinct images. Classification of this validation set of into the 2 posture categories achieves a maximum accuracy of 91.329%.



	Name	Smoothed	Value	Step	Relative
●	Standing Correct vs Incorrect	0.06111	0.07699	1.290K	9m 56s
●	Sitting Correct vs Incorrect	0.05055	0.04299	1.420K	11m 22s

Fig. 10. Behaviour of Validation Loss vs Number of Iterations

The validation loss decreases quickly initially and appears to plateau around 0.09 after 500 iterations for standing posture and the validation loss appears to plateau around 0.06 after 300 iterations for sitting posture as shown in Fig.10. The validation loss can be further decreased by fine-tuning the hyper parameters of the model (learning rate, step size, regularization strength) and reducing the complexity of the model.



	Name	Smoothed	Value	Step	Relative
●	Standing Correct vs Incorrect	0.9816	0.9770	1.290K	9m 56s
●	Sitting Correct vs Incorrect	0.9736	0.9750	1.420K	11m 22s

Fig. 11. Accuracy for Posture Classification with Number of Iterations

Similarly, the training accuracy increases steeply and plateaued around 1,000 iterations for standing posture, while the training accuracy shows only a slight increase for the case of sitting posture as shown in Fig.11. Further the simulation was carried out by varying the number of CNN layers, as shown in Fig.12. The accuracy is highly dependent on the number of CNN layers. Accuracy increases significantly on increasing the number of CNN layers. However, after a particular value, the accuracy starts to show no change. It is therefore observed that the maximum value of accuracy, 91.329% is obtained while using 4 CNN layers.

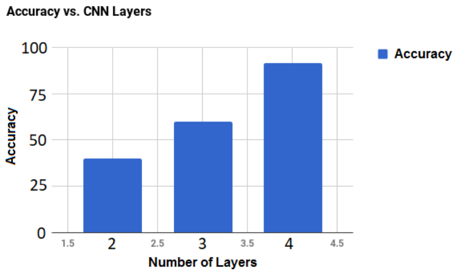


Fig. 12. Training Accuracy Results with regard to Number of CNN Layers

IV. Conclusion

The above results suggest what a promising tool CNNs can be, especially to solve visual recognition tasks. With such competitive simulation time and easy implementation, CNNs have an added advantage of their holistic construct which offers local connectivity and parameter sharing which makes them all the more a preferred choice for such Computer Vision tasks. The present work shows how postures can be classified by modelling the problem with a CNN. The work's application of CNN on posture classification achieves results to the extent of 91.3% accuracy on real time environment situations with a simplified approach.

V. Future Work

In future, we plan to extend the proposed model to classify the types of incorrect postures like rounded shoulders, forward neck and hunched back for both standing and sitting postures. We also would like to experiment with a significantly larger dataset by taking the full body posture into consideration.

Acknowledgment

The authors would like to acknowledge the authorities of BITS Pilani, Dubai Campus for their constant support and motivation. The authors would also like to thank Dr. Prachi Jain, Senior Physiotherapist for her valuable suggestions.

References

[1] Samreen Yasmeen, Muhammad Shahrukh and Zuhaira Farooqui, "Postural Awareness In School Going Students And Teachers," *Pakistan Journal Of Rehabilitation*, vol. 3, no. 1, pp. 39-45, 2014.

[2] Adila Md Hashim and Siti Zawiah Md Dawal, "Comparison of Working Postures among Students in School Workshop: A CAD Environment Analysis," *Advanced Engineering Forum*, vol. 10, pp. 199-206, 2013.

[3] Michal Latalski, Jerzy Bylina, Marek Fatyga and Tadeusz Trzpis, “Risk factors of postural defects in children at school ages,” *Annals of Agricultural and Environmental Medicine (AAEM)*, vol. 20, no. 3, pp. 583-587, 2013.

[4] Esther Liyanage, Indrajith Liyanage and Masih Khan, “Efficacy of Isometric Neck exercises and stretching with ergonomics over ergonomics alone in Computer Professionals,” *International Journal of Scientific and Research Publications (IJSRP)*, vol. 4, no. 9, Sep. 2014.

[5] Wang C, Tian F, Zhou Y, He W and Cai Z, “The incidence of cervical spondylosis decreases with aging in the elderly, and increases with aging in the young and adult population: a hospitalbased clinical analysis,” *Clinical Interventions in Aging*, vol. 11, no. 11, pp. 47-53, 2016.

[6] Max Jaderberg, Karen Simonyan, Andrew Zisserman and Koray Kavukcuoglu, “Spatial Transformer Networks,” *28th International Conference on Neural Information Processing Systemss (NIPS’15)*, pp. 2017-2025, 2015.

[7] Andrej Karpathy, “Convolutional Neural Networks for Visual Recognition,” in *Lecture Notes in Stanford CS class CS213n 2016*.

[8] Praveen Kumar B.M and Suma Huddar, “A Background Subtraction Technique For Object Detection Using SVM,” *International Journal Of Advanced Computing And Electronics Technology*, vol. 3, no. 2, pp. 2394-3408, 2016.

[9] Swapnil V Tathe and Sandipan P Narote, “Face Recognition in Videos using Gabor Filter,” *IOSR Journal of Computer Engineering (IOSR-JCE)*, pp. 75-81, 2017.

[10] Bohyung Han and Larry S. Davis, “Density-Based Multifeature Background Subtraction with Support Vector Machine,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 5, pp. 1017-1023, 2012.

[11] Thierry Bouwmans, Fatih Porikli, Benjamin Hferlin, and Antoine Vacavant, *Background Modeling and Foreground Detection for Video Surveillance*, 1st ed. Taylor & Francis Group, 2015.

[12] Jianxin Wu, *Introduction to Convolutional Neural Networks*, LAMDA Group National Key Lab for Novel Software Technology Nanjing University, China, 2017.

[13] Samer Hijazi, Rishi Kumar, and Chris Rowen, *Using Convolutional Neural Networks for Image Recognition*, IP Group, Cadence, 2015.

[14] Yangqing Jia and Evan Shelhamer, *Caffe - Deep Learning Framework Notes*, Berkeley AI Research(BAIR).

[15] Wunpen Chansirinukor , Dianne Wilson, Karen Grimmer and Brenton Dansie, “Effects of backpacks on students: Measurement of cervical and shoulder posture,” *Australian Journal of Physiotherapy*, vol. 47, pp. 110-116, 2001.