International Journal of Science and Research (IJSR)
ISSN: 2319-7064
ResearchGate Impact Factor (2018): 0.28 | SJIF (2018): 7.426

3D Try on Using Deep Neural Networks

Sai Medavarapu

Texas A&M University - Corpus Christi
Email: *smedavarapu1[at]islander.tamucc.edu*

Abstract: ***This paper proposes a virtual try on system where user can try the eye glasses on the 2D face in real time. The fundamental specialized test of this framework is the programmed 3D eyeglasses model remaking from the 2D glasses on a frontal human face. This paper initially proposes a convolutional neural network model to label the facial key points. Generated model is used for placing the glasses on the detected facial key points. The test results exhibit the efficiency on execution of this methodology and the business capability of proposed methodology. Code is made available at https://gitlab.tamucc.edu/smedavarapu1/ase-capstone.git***

Keywords: Facial landmark detection, Human Activity Recognition, Masked object detection, contour detection, 3D Try on

1. Introduction

Virtual Try on stand out amongst most significant realistic applications for humans. It gives a methodology that is very intuitive to assess structures of items, for example, attire, footwear, etc. Since, virtual try on provides a convenient and effective way to try eye glasses with respect to personal taste and choice, it has recently caught attention of business applications [9].

Facial landmark detection [14] aims to detect the location of facial landmarks, such as the corners, center of the eyes, eyebrows, the tip of the nose. It has drawn much attention recently as it is a prerequisite in many computer vision applications. For example, facial landmark detection can be applied to a large variety of tasks, including face recognition, head pose estimation, facial reenactment and 3D face reconstruction etc. In the existing system [7] the image is passed to the model and it is processed for a while. After processing, it lets the user to try the eye glasses on the face. However, user may be interested in trying the multiple glasses in real time. This is not addressed by the existing system. Fig 1 demonstrates the detection of the facial landmarks when the input is passed and key points viz., center of the eye, eye corners etc are detected.

Figure 1: Input and output of facial landmark detection, left image is the input and right image is the detected landmarks on the face

In brief time frame, Faster RCNN and Fully Convolutional networks (FCN) have gained lot of importance in computer vision applications, for example, object detection and semantic segmentation etc. Flexibility and robustness are offered by these two methods in training and inference time. Faster RCNN is extended to Mask-RCNN [8] constructs the branch properly on the object input and output for exact boundaries. Contour detection serves as the basis of a variety of computer vision tasks such as image segmentation and object recognition. Various computer vision tasks are served by contour detection using image segmentation and object detection. Edge detection is less challenging when compared to contour detection. According to Martins [4] definition, Edge detection requires the particular color, brightness and texture. Alternatively, contour detection focuses on pixel detection using the object contours. The basic procedure for contour detection is done using the image pixels and the image is passed through a classifier to determine the contour. Hand designed feature [3] was previously used in the past decade where the contours where detected using it. In low level image settings, the previous methods have issue in detecting the corners of the image. After the advent of the deep learning, many researchers have focused on improving the hand deigned features by using the SIFT [1] and HOG [5].

The motivation behind this area of interest is in today's world 7 out of 10 are using glasses. Figuring out the good eyeglass frame was arduous task for me since childhood. It not only wasted the time but also did not give me the right choice. This paper proposes a intuitive eyeglasses virtual try on method, which takes the input of glasses from the database. The user should sit in front of the web camera and The customer can visualize the augmented video stream and interact with the virtual eyeglasses by moving and rotating her/his head. If the user want to try the other glass it can be achieved by interacting with the next frame button. The glass can be changed by using the masked object [13]. Following are the few key contributions that paper makes.

- Facial landmark detection, The facial landmarks can be detected by rough pixel parsing step using a convolutional neural network (CNN).
- Detecting the contours using deep learning.
- Given the detected facial landmark model, this virtual try on methodology uses a web camera to try the glasses in real time.

The rest of the paper is organized as follows, Section II describes the literature work and the background review done in order to achieve this. Section III Describes the existing system and technical details of it. Section IV represents the technical details including the architecture of it. Section V Explains the Experimentation, results and future work. Followed by conclusion in the later section.

2. Related Work

I reviewed prior works from few perspectives viz., Facial landmark detection, Human Activity recognition, Masked Object Detection, Contour Detection.

a) 3D Object detection

In the work proposed by Michael Danielczuk et al [12] The ability to segment unknown objects in depth images has potential to enhance robot skills in grasping and object tracking. Recent computer vision research has demonstrated that Mask R-CNN can be trained to segment specic categories of objects in RGB images when massive hand-labeled datasets are available. As generating these datasets is time- consuming, they instead train with synthetic depth images. Many robots now use depth sensors, and recent results suggest training on synthetic depth data can transfer successfully to the real world. they present a method for automated data set generation and rapidly generate a synthetic training data set of 50,000 depth images and 320,000 object masks using simulated heaps of 3D CAD models. They train a variant of Mask R-CNN with domain randomization on the generated dataset to perform category-agnostic instance segmentation without any hand-labeled data and we evaluate the trained network, which they refer to as Synthetic Depth (SD) Mask R-CNN, on a set of real, high-resolution depth images of challenging, densely-cluttered bins containing objects with highly-varied geometry. In the work proposed by Akshay, [1] This lets the user detect the facial key points on the face. Where user can try the sun glasses.

b) Facial Landmark Detection:

Facial landmark detection by using GANs [10]. Here, they proposed a style aggregated approach to deal with the large intrinsic variance of image styles for facial landmark detection. The goal of facial landmark detection is to detect key points in human faces, e.g., the tip of the nose, eyebrows, the eye corner and the mouth. Facial landmark detection is a prerequisite for a variety of Deep Neural Network applications. Yu Chein et al proposed a methodology [7] Face Super-Resolution (SR), a.k.a. face hallucination, aims to generate a High-Resolution (HR) face image from a Low-Resolution (LR) input. It is a fundamental problem in face analysis, which can greatly facilitate face-related tasks. This methodology makes the low-resolution images as high resolution with more clarity using Face Super Resolution Generative Adversarial Network (FSRGAN) to incorporate the adversarial loss into FSRNet. Plotting the convolutional neural network architecture is the arduous task sometimes.

In the work given by Harish Iqbal[2], he proposed a latex code to draw the architecture.

c) Human Activity Recognition

While facial landmarks let the neural network know about the landmarks. Human facial movements play crucial role in accurate detection. Fabien baradel et all. In [14] described a methodology for human activity recognition by using the RGB data which does not rely on any pose during test time. They used the Recurrent Neural Networks. To demonstrate the activity, they used two workers These workers receive the glimpses, jointly performing subsequent motion tracking and prediction of the activity itself.

d) Masked Object Detection:

Masking allows us to handle variable length inputs in RNNs. Although RNNs can handle variable length inputs, they still need xed length inputs. Therefore, what we do is to create a mask per sample initialized with 0 with a length equal to the longest sequence in the dataset. Xiangyun Zhao et al [6], proposed a methodology which augments the object detector with generated object masks from the bounding box annotation, named as Pesudo-mask Augmented Detection (PAD). It starts from a strong baseline network architecture that directly integrates the state-of-the-art Fast-RCNN network for object detection and InstanceFCN for object segmentation, in a normal multi-task setting.

e) Contour Detection

Contour detection [11] serves as the basis of a variety of tasks such as image segmentation and object recognition in Deep Convolutional Neural Networks. Though there are automated approaches for the contour detection [2], However, even the latest evolution's struggle to precisely delineating borders, which often leads to geometric distortions and inadvertent fusion of adjacent building instances. Marcos et al. proposed a methodology [13] to learn Active Counter Models parameterization using CNN Convolutional Neural Networks which has 7 convolutions with 3 * 3 size followed by ReLu and Max-pooling

f) Deep learning

In computer vision- based applications, deep learning has achieved a great success in areas like Image recognition and object detection. In the paper proposed by Wei Shen et al [11], they explain the importance of deep learning in training the supervised application by using the trick Rectified Linear Unit(ReLu) activation function.

g) Virtual Try On

In the work proposed by Jianzhu et al [7] in the appli- cation of face recognition. They estimate position of the 3D points on the face and place the glasses on the passed image. The 3DMM fitting model [15] explains a methodology on extracting the 3D faces from the passed image. This methodology is used by zhu et al [7] in extracting the 3D

[1] https://towardsdatascience.com/73

[2] https://github.com/HarisIqbal88/PlotNeuralNet

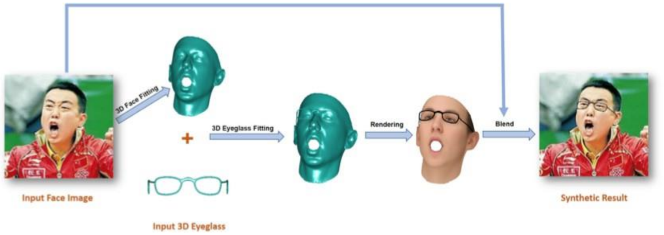


Figure 2

face from the image. Once the image has been extracted, they key points are detected and placed on the generated 3D face as shown in figure. The Zmodel and phong illumination model algorithm are used to generate the original image by blending the generated model.

The 3D eyeglass tting problem is formed as Eq. 1, where f is the scale factor, Pr is the orthographic projection matrix, p g is the anchor points on 3D eyeglass, p f is the anchor points on reconstructed 3D face model, R is the 3 3 rotation matrix and t, 3d is the translation vector.

$$\arg\ \min_{f,Pr,R,3d} \left\| f \cdot P_r \cdot (P_g + t_{3d}) - p_g \right\| \quad (1)$$

3. Proposal

a) *Dataset Description*

The dataset for this project was used from Kaggle [3] which was provided by the Dr. Joshua[4] of the university of Montreal. Each predicted keypoint is specified by an (x,y) real-valued pair in the space of pixel indices. There are 15 key-points, which represent the following elements of the face:

Left eye center, right eye center, left eye inner corner, left eye outer corner, right eye inner corner, right eye outer corner, left eyebrow inner end, left eyebrow outer end, right eyebrow inner end, right eyebrow outer end, nose tip, mouth left corner, mouth right corner, mouth center top lip, mouth center bottom lip

Left and right here refers to the point of view of the subject. In some examples, some of the target key-point positions are missing (encoded as missing entries in the csv, i.e., with nothing between two commas).

The input image is given in the last field of the data files, and consists of a list of pixels (ordered by row), as integers in (0,255). The images are 96x96 pixels.

training.csv: list of training 7049 images. Each row contains the (x,y) coordinates for 15 keypoints, and image data as row-ordered list of pixels. test.csv: list of 1783 test images. Each row contains ImageId and image data as row-ordered list of pixels submissionFileFormat.csv: list of 27124 key- points to predict. Each row contains a Row Id, Image Id, Feature Name, Location. Feature Name are ”left Eye Center x”,” ”right Eyebrow Outer endy,” etc. Location is what you need to predict.

[3] https://www.kaggle.com/c/facial-keypoints-detection/overview/description

[4] https://mila.quebec/en/yoshua-bengio/

b) *System Overview*

This paper proposes a intuitive eyeglasses virtual try on method, which takes the input of glasses from the database. The user should sit in front of the web camera and the customer can visualize the augmented video stream and interact with the virtual eyeglasses by moving and rotating her/his head. If the user wants to try the other glass it can be achieved by interacting with the next frame button. The glass can be changed by using the masked object [13]. This system runs on a computer with an Intel Xeon CPU @ 2.50
GHz and 8-GB RAM. In the try-on module, a web camera is equipped to capture users color and depth streams.

c) *Finding the contours*

The shape of eyeglasses has horizontal symmetry intrinsi- cally. As for inner contours, the left and right two contours are symmetric, and as for the outer contour, the left and right parts are symmetric. However, the input image is often not strictly horizontal and the human head in the image usually has a little yaw, pitch, and roll movements from a neutral pose. The metric evaluation of the pairs can be calculated by using the formula.

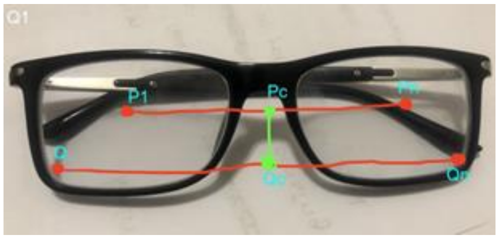


Figure 3: Finding Contours by using the metric evaluation

$$f(Q_i + Q_n) = \left\| (P_n - P_1) \cdot (P_c - Q_c) \right\| \quad (2)$$

P1 is the left Eye center, Pn is the Right eye center, Q1 is the left Eye corner, Qn is the right eye corner. Pc and Qc are the centers of the frame which are the tip of the nose and center between two eyes.

4. Architecture

The architecture of the entire system is represented in the figure 4. It is the convolutional neural network with 6

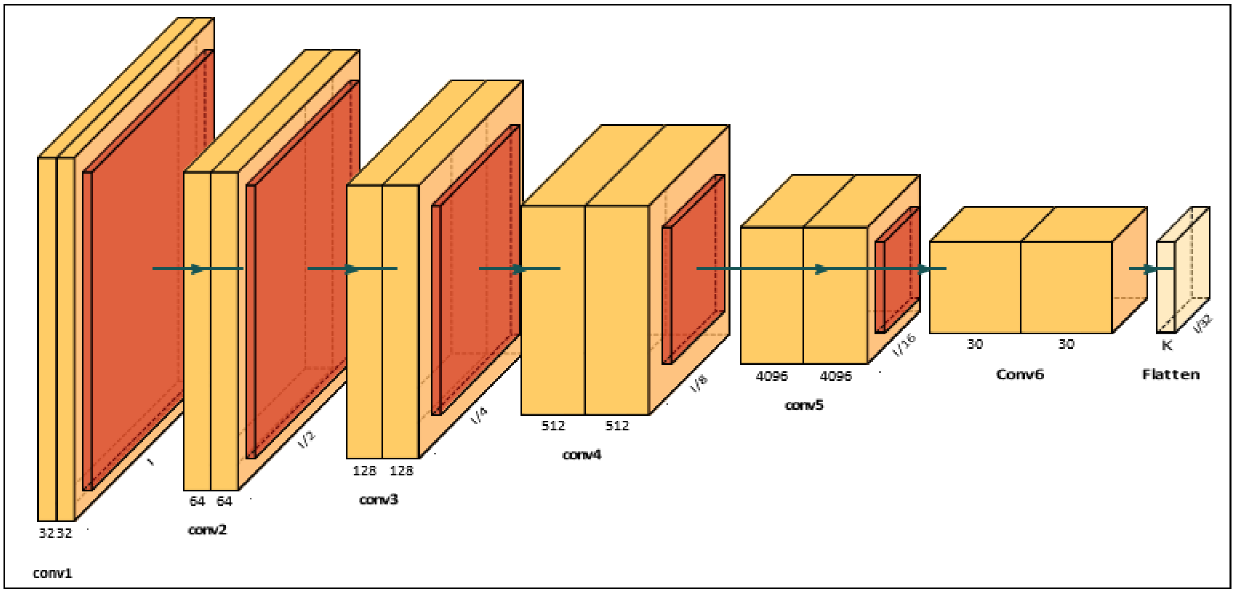


Figure 4: CNN Architecture of the system

convolution layers. In the first convolution, the input image of size 96 grey scale is passed to the layer with size 32 with activation function ReLU(Rectified Linear Unit) and later Maxpooling with size 2 is performed. The generated image is passed to the next convolution of size 64 which has the size 3 and Rectified Linear Unit activation function. Later, the Maxpooling of stride 2 is performed and the dropout later is added.

In the next layer, the image is passed into size 128 with the softmax activation function is processed and maxpooling with stride 3 is performed and dropout layer is added. The next layer in the architecture is of size 512 which has the maxpooling with stride 2. Since, there are 15 key pair of points in the available data. The next convolution with size 30 is added and the activation function ReLU is added to it and maxpooling with stride 2 is processed. The next step is to add flatten. In the later steps, the dense layers with size 64, 128, 256, 64 are added to get the output of detected eye points on the face.

5. Experimentation and Results

a) Experimentation

All face images are resized to 96 size. The mean squared error loss function and adam optimizer is used to optimize the network. Based on these configurations, the training data can process the data as 10 images per second in normal CPU settings of cores. When tried on the GPU, it can process upto 200 images per second. The user should sit in front of the web camera and the customer can visualize the augmented video stream and interact with the virtual eyeglasses by moving and rotating her/his head. If the user wants to try the other glass it can be achieved by interacting with the next frame button. The glass can be changed by using the masked object [13]. This system runs on a computer with an Intel Xeon CPU @ 2.50 GHz and 8-GB RAM. In the try-on module, a web camera is equipped to capture users color and depth streams.

b) Results

This project was tested on the two different datasets 1) Kaggle facial landmark data set 2) MS celeb data set for comparison between two approaches. In the first data set which is Kaggle facial landmark dataset the data was run for 100 epochs using adam optimizer function. The accuracy when Kaggle dataset was used is 78% and for the existing system the percentage was 83%. When the data of the MSceleb was used the results were 88% and the existing system results were 96% . Finally, The plots between the values for accuracy was plotted for last 10 epochs over 100 epochs for the proposed system and the loss function values were also plotted for the proposed system. The figure 5 explains the plot accuracy for the proposed system when run on the Kaggle facial landmark data set.

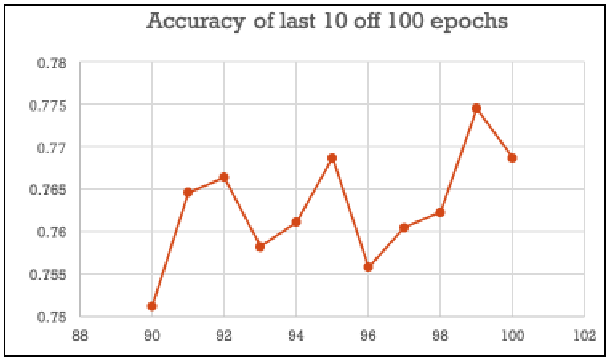


Figure 5: Accuracy for the proposed system for the last 10 epochs

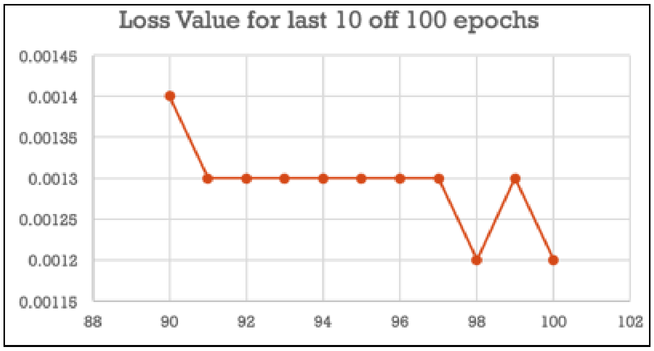


Figure 6: Loss values for the proposed system on Kaggle facial landmark dataset.

Figure 6 demonstrates the loss value of last 10 epochs of system when run on 100 epochs. The loss value of the function was 0.0012 for the last epoch. The dataset used for this was the Kaggle facial landmark dataset.

Table I: Comparison on Kaggle facial landmark dataset

	Accuracy	Loss Value
Existing Approach [7]	85%	0.00017
Proposed Approach	78%	0.0012

Table II: Comparison on MSCeleb dataset

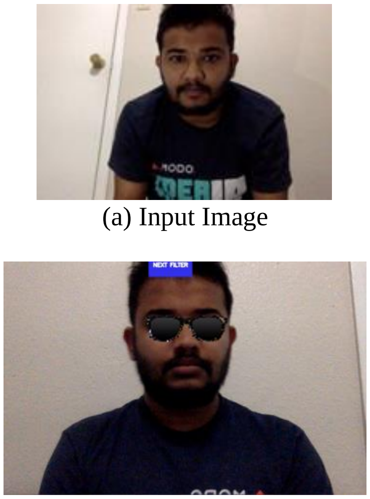
	Accuracy	Loss Value
Existing Approach [7]	96.61%	0.0007
Proposed Approach	87%	0.0015

For comparison of the results, the two datasets were trained on GPU for 100 epochs and the Table I, Table II are the comparison based demonstration of the proposed and the existing approaches. The Table I is the demonstration of the accuracy and the loss value for the kaggle data set. The accuracy when the existing system was tried with the changes in the code with respect to the dataset labeling of the proposed system us 85% and the proposed system accuracy was 78%. Table II represents the accuracy and loss values with respect to the MSCeleb dataset. Since the code for existing approach was available for MSCeleb dataset. The proposed system was executed on the MSCeleb dataset with changes in code with respect to the MSCeleb dataset labeling. The accuracy of the existing system was 96% and for the proposed system the accuracy is 87%

The figure 7 explain when the image is passed from the webcam using the tensorflow library it is processed from the architecture and the detected facial landmarks can be used to try the glasses on the image. If the user want to check the other glasses they can click the next glass button to change it to the other glass as shown in figure.

6. Conclusion and Future work

This system provides the novel methodology for trying the glasses in real time using the tensorflow library and Deep learning. This system can be used when the user want to try the glasses from home. Results from the experiments show that they are satisfactory and give the virtual experience of the try on and letting users to manage their looks more conveniently. The limitation of the system is that it can be tried on when the methodology is able to detect the inner contour and outer contour of the glasses. The future work of the system can be extended by using more training data to detect the key points and improving the methodology for more accurate contour detection.



(a) Input Image

(b) Output image with glasses

Figure 7: Try on result of the system

References

[1] G. Lowe. ;. Distinctive image features from scale-invariant keypoints. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*, 2004.

[2] Julien Mille Graham W. Taylor: abienBaradel, Christian Wolf. Glimpse clouds: Human activity recognition from unstructured feature points. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 469-478.*, 2018.

[3] S. Guadarrama Contour shick and T. Darrell. Caffe: ;. Convolutional detection and hierarchical image segmentation. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*, 2011.

[4] Fowlkes D. R. Martin and J. Malik.;. Learning boundaries using color, brightness. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*, 2004.

[5] N. Dalal and B. Triggs. ;. Histograms of oriented gradients for human detection. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*, 2005.

[6] De Xu et al. Hu Su. An overview of contour detection approaches. *The International journal for automation and computing December 2018.*, 2018.

[7] Zhen Lei Jianzhu Guo, Xiangyu Zhu ? and Stan Z. Li;. Face synthesis for eyeglass-robust face recognition,. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*, 2018.

[8] J. Donahue; Kaiming He. Mask r-cnn. *The IEEE Conference on Computer Vision and Pattern Recognition*

(CVPR)., 2018.

[9] Farbiz A. Niswar M. Yuan, I. R. Khan and Z. Huang;. Fa mixed reality system for virtual glasses try-on,. *in Proc. 10th Int. Conf. Virtual Reality Continuum Appl. Ind, pp. 363366.*, 2011.

[10] Diego Marcos, Devis Tuia, Benjamin Kellenberger, Lisa Zhang, Min Bai, Renjie Liao, and Raquel. Urtasun. . learning deep structured active contours end-to-end. *The IEEEConference on Computer Vision and Pattern Recognition (CVPR) June 2018.*, 2018.

[11] Ahmed Khattab Marwa Mamdouh, Mohamed A. I. Elrukhsi. Deep- contour: A deep convolutional feature learned by positive-sharing loss for contourdetection. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.A)*, 2016.

[12] Saurabh Gupta Andrew Li Andrew Lee Jeffrey Mahler Ken Goldberg ; Michael Danielczuk, Matthew Matl. Segmenting unknown 3d objects from real depth images using mask r-cnn trained on synthetic data. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*, 2019.

[13] Yichen Wei Xiangyun Zhao, Shuang Liang. Pseudo mask augmented object detection. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.*, 2018.

[14] Wanli Ouyang Yi Yang: Xuanyi Dong, Yan Yan. Style aggregated network for facial landmark detection. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 379-388.*, 2018.

[15] Yan J Yi D Li SZ ; Zhu X, Lei Z. High-delity pose and expression normalization for face recognition in the wild. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*, 2016.

DOI: https://dx.doi.org/10.21275/SR24801075537