Specimen 'A': Title Sheet

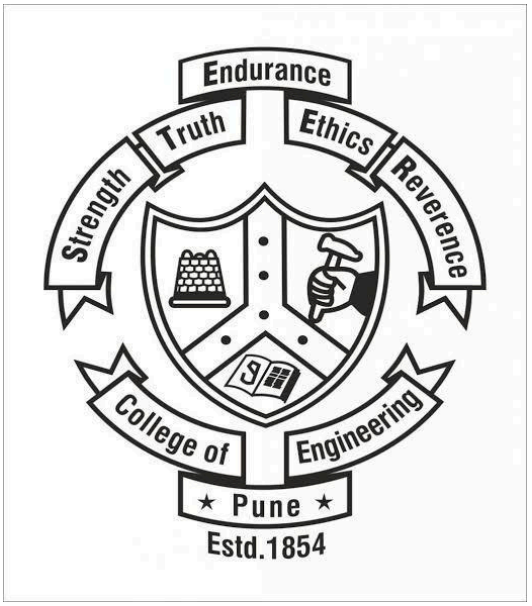
An Intelligent Pose Recommendation System using feature fusion for humans

Submitted in fulfilment of the requirements
of the degree of
Bachelor of Technology

By

Sakshi Patil (111807032)
Piyusha Taware (111807039)
Sonali Sargar (111807053)

Guide (s):
Dr. Abhishek Bhatt



Electronics and Telecommunication Engineering
COLLEGE OF ENGINEERING PUNE
April, 2022

CONTENTS

List of Figures:

Abbreviations:

AP: Average Precision

CNN: Convolutional Neural Network

COCO: Common Objects in Context

DARKPose: Distribution-Aware Coordinate Representation for Human Pose Estimation

HRNet: Deep High-Resolution Representation Learning for Human Pose Estimation

HSV: Hue Saturation Value

mAP: mean Average Precision

MPII: Max Planck Institute for Informatics

NasNet: Neural Architecture Search Network

PCKh: Head-normalized Probability of Correct Keypoint

ResNet: Residual Neural Network

VGG: Visual Geometry Group

1. ABSTRACT

With rapid development of technology and easy access to mobile phones and other smart devices, there has been an increase in the number of social media platforms.
Since the pandemic, globally a hike in the number of social media users is observed. It has become a virtual platform for people to express themselves. To gain more attention and audience, people want to have the best photos and project themselves in a better way. However, it is hard to guarantee the quality of such uploaded images because of poor image acquisition devices and lack of professionalism in the pictures as everyone is not a professional photographer. In order to tackle these problems, this study proposes an intelligent photo pose recommendation method to recommend professional photo pose according to the background of the input image. Given a query image with only background, the global features are extracted from it using VGG 16 and ResNet 101. Secondly, the local features are extracted using colour histograms. These features are concatenated to give a final feature vector. The feature vector then is used to get the most similar image. Open pose is used to extract pose from the most similar image and is superimposed on the query image to give pose recommendation.

2. KEYWORDS

Convolutional Neural Networks: ResNet101, ResNet152, VGG16, Vgg19, Evaluation metrics: Chi Square Distance, Cosine Similarity, Euclidean Distance, Human Pose Estimation, Human Pose Recommendation.

3. INTRODUCTION

Human pose understanding has big scope with respect to increasing the aesthetic of photographs and AI based action triggers. Sophisticated systems to identify nuance of human poses open up opportunities in Camera differentiation. In an era where almost everyone owns a mobile phone and most of the pictures taken end up in social media, posing your subject provides the opportunity to tell a different story with every frame. At the same time, it helps improve the appearance of the subject in the photo. The difference between professional and unprofessional photos can be easily spotted. For example, the pictures in *Figure 1* show unprofessional photos and *Figure 2* show professional photos

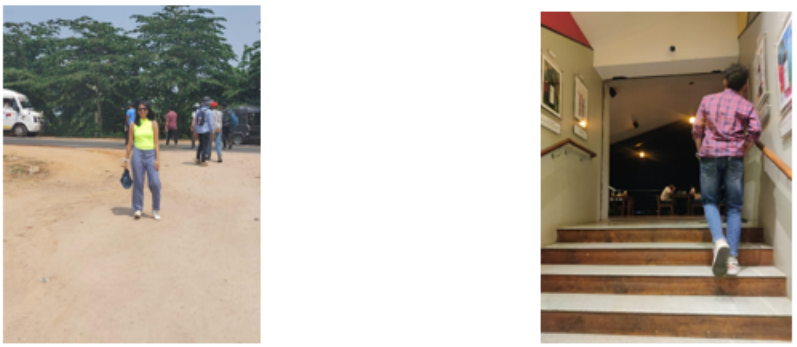


Figure 1:Unprofessional Photo

Figure 2: Professional photos

The approach used in this paper first extracts the global features using concatenation of features from VGG16 and ResNet101. The global features capture the context of the entire image. Secondly color histogram is used for extraction of local features. Region based color histograms represent the color distribution in the specified regions. The local and global features are concatenated to give a final representation of the query image. The feature vector is calculated for all images in the dataset as well. Cosine similarity is used for finding the most similar image from the dataset. Once the most similar image is found, Open pose is used to extract the pose of the person from the image. This pose is superimposed on the query image in the form of a skeleton.The dataset used for the study was collected manually from

professional photography websites. Experiments and user study help in evaluating the efficiency of the approach.

4. LITERATURE REVIEW

4.1 Pose Estimation

Open Pose [7] was proposed by Zhe Cao et. al. in 2019. It is a bottom up approach for multi person pose estimation. The aim of the research is overcoming the difficulties faced while detecting individual body joints in multi-person images.
Given an input image, the location of each joint is determined by part confidence maps, the orientation of the body parts is determined by a 2D vector that represents the degree of association between the body parts known as Part Affinity Fields. bipartite matching is performed on these body part candidates. Weaker links in the bipartite graphs are pruned using the Part Affinity Fields values and finally, human pose skeletons are estimated and assigned to every person in the image.
The architecture has a two-branch and multi-stage CNN network. The input to it is a feature map F of an image which is initialised by the first 10 layers of VGG architecture. The model simultaneously predicts confidence maps S for predicting location of joints and affinity fields L for depicting association between body parts.
For next stages, the inputs will be previous stage confidence maps, part affinity fields and feature map F. By bipartite matching, 2D key points for every individual in the image are obtained. The evaluation of this work is done on COCO and MPII datasets using AP, mAP and PCKh@0.5 evaluation metrics to achieve best results compared to existing state of the art models.A high accuracy and real time performance is achieved, regardless of the number of people in the image.
However it is not very successful for non typical poses and upside-down examples.

HRNet - Deep High-Resolution Representation Learning for Human Pose Estimation. It [8] proposed a model to maintain high resolution representation throughout the entire process, unlike the usual trend of first downsampling and then recovering high resolution values. The architecture begins with a high-resolution subnetwork as the first stage, high-to-low resolution subnetworks are gradually added one after another in parallel to form more stages. HRNet architecture benefits in obtaining more precise predicted heatmaps and more accurate spatial resolution. Multi-scale fusions are conducted repeatedly for information exchange across subnetworks. Hence, use of intermediate heatmap supervision is eliminated. During forward propagation of each scale branch the resolution of the feature maps is unchanged.Although information is exchanged between each scale branch, all branches are different.
The performance of HRNet is evaluated on COCO and MPII datasets using AP, mAP and PCKH@0.5 evaluation metrics. HRNet can also be used effectively for video pose tracking on PoseTrack dataset.

DARKPose - Distribution-Aware Coordinate Representation for Human Pose Estimation. It [9], focuses on the coordinate representation of heatmaps. In DARKPose, predicted heat maps are decoded into final joint coordinates in the original image. Motivation behind DARKPose is: 1) Efficient coordinate decoding based on Taylor-expansion , and (2) unbiased sub-pixel centred coordinate encoding. The DARKPose method effectively allows the use of small resolution input images with slight performance

degradation, while substantially increasing inference efficiency of the model hence promoting embedded AI low latency and low-energy applications.
Heatmap distribution is modulated before resolution upsampling to reduce the effect of multiple peaks around maximum activation and reduces performance degradation of the decoding method. Coordinate decoding method employs three steps: 1) Heatmap distribution modulation, 2) Distribution aware joint localization using Taylor expansion as sub pixel accuracy, and 3) Resolution recovery to the original coordinate space. Coordinate encoding also encounters the limitation of resolution reduction. Hence, the encoding method starts by downsampling the provided original image to model input size. Thus, ground truth join coordinates need to be transformed accordingly before heatmap generation to ensure unbiased sub-pixel centered coordinate encoding.
The performance of the model is evaluated on MPII and COCO datasets using PCK and OKS measure respectively.

4.2 Pose Recommendation

In **Adaptive recommendation for photo pose via deep learning** [10], Tong Hao1 et. al. propose fusion of global and local features for pose recommendation. First, the Global features are extracted from each photo with CNN model, VGG16. Then Salient Region Detection is used to find the Region Of Interest, it is utilized to find important regions in each photo and then Histogram of oriented Gradients (HoG) is used to represent local structural information. Feature vectors are obtained after this step. The next step is feature fusion.CCA (Canonical Correlation Analysis) is used to utilize both local saliency and global context for diverse feature fusion.
Lastly, Euclidean distance is calculated to handle similarity between user's photos and professional photos.

Intelligent Portrait Composition Assistance[11] proposes an intelligent framework of portrait composition using deep-learned models and image retrieval methods. Specifically, it addresses aesthetic retrieval and evaluation of the human poses in portrait photography, and tries to improve the quality of the next shot by providing meaningful and constructive feedback to an amateur photographer. A highly-rated web-crawled portrait dataset is exploited for retrieval purposes. This framework detects and extracts ingredients of a given scene representing a correlated hierarchical model. It then matches extracted semantics with the dataset of aesthetically composed photos to investigate a ranked list of photography ideas, and gradually optimizes the human pose and other artistic aspects of the composed scene supposed to be captured.

Aesthetic Composition Representation for Portrait Photographing[12], this paper presents an intelligent portrait photographing framework for automatically recommending the suitable positions and poses in the scene of photography taken by amateurs. By analyzing aesthetic characteristics features, it proposes a solution by constructing aesthetic composition representation which covers the attention composition and geometry composition to identify the underlying technique of a professional photographer. First, it extracts the attention composition feature of the professional photo by utilizing a visual saliency model. Then, a geometry composition feature is also presented to learn the spatial correlation. Finally, composition rules are applied to make appropriate pose and position. To represent scene structure without stressing the portrait, it designs a decaying exponential function to weaken the magnitude of saliency while preserving spatial attention distribution. To find the

nearest-neighbors for the query images in the quantized 1024-dimension low level visual feature space Hierarchical Kmeans method is used to speed up the search process.

5. DATASET

As photo pose recommendation is a relatively newer concept, a lot of work has not been done in these areas and so there is no specific dataset available. Previous works on this topic collect private dataset to evaluate their performance. So, we created our own dataset of 1067 images. We looked for professional photography websites for free downloading of photos. Since manually downloading photographs from the website is time consuming, we wrote a python script for crawling photos from a professional photo website, namely Getty Images and StockSnap. It contains photos from millions of creative photographers around the world. The website has photographs belonging to different categories like food, technology, nature etc. Since our topic is human pose recommendation, we selectively crawled images which belonged to people, kids, women, men, and family categories. We managed to collect around 1067 images (Can be further extended). We resize the image to 224*224 size for input to VGG16 model. The image is converted to an array and mean RGB intensity is subtracted from it. Our dataset contains images from different backgrounds such as mountains, beach, office, building, garden, bench, etc. Hence, our dataset contains most of the common scenarios and can be utilized for recommending the images for various backgrounds.

6. METHODOLOGY

Our approach aims at recommending a photo pose which is best suited for the input background image. This approach tries to compare the query image with our dataset and finds the most similar image based on cosine similarity. Framework of our approach consists of 5 steps as shown in *Figure 4*: 1) extracting global features: CNN models VGG16 and ResNet101 are implemented separately to extract global features of given image which gives context information of the input, 2) Feature concatenation: feature vectors obtained from VGG16 and ResNet101 are concatenated to form one feature vector for further processing, 3) extracting local features: color histogram is used to get color distribution of the image, 4) feature fusion: obtained global and local features are concatenated to get better representation of images, 5) similarity measure: cosine similarity metric is used to evaluate similarity between query image and images from dataset.

Global features extraction and concatenation:
To extract the global features of a given image using CNN we stop propagation at an arbitrary, but pre-specified layer (such as an activation or pooling layer). Then, extract the values from the specified layer and treat the values as a feature vector.
In the presented method,a hybrid model of VGG16 and ResNet101 is used to extract global features. VGG16 outputs a feature vector of 4096 dimension, and ResNet101 outputs a feature vector of 2048 dimension. These two feature vectors are then concatenated to obtain a combined global feature vector of 6144 dimension.

Local features extraction:
To extract local features of a given image we use color histogram in the following way:
For the image descriptor, we divide our image into five different regions as shown in *Figure 3*: (1) the top-left corner, (2) the top-right corner, (3) the bottom-right corner, (4) the bottom-left corner, (5) the center of the image. As shown in the image below.

Figure 3: Division of a image into local regions

By using these identified regions we'll be able to imitate a crude form of localization, and so we can represent our above beach image as a combination of blue sky in the top-left and top-right corners, brown sand in the bottom-left and bottom-right corners, and then a mixture of blue sky and brown sand in the center region.
The 3D color histogram in the HSV color space has 12 bins for the saturation channel, 8 bins for the Hue channel and 3 bins for the value channel. So, a 3D color histogram yields a feature vector of 288(*12 x 8 x 3)* dimensions for one region. Hence, we obtain a local feature vector of dimension 1440, considering all the 5 identified regions.

Feature fusion and similarity measure:
To obtain better representation of an image, we concatenate the obtained global and local features and form a combined feature vector of dimension 7584. This feature vector now contains both the context information and color histogram based local information of images. We obtain fused feature vectors for all images in the entire dataset and one for the query image. Then the feature vector of the query image is compared with the feature vectors of the images in the dataset using cosine similarity metric and similarity score of the query image and all the images from the dataset are evaluated.
The cosine similarity metric can be formulated as:

$$similarity = cos(\Theta) = S(I_q, I_{id}) = \frac{\sum_{i=1}^{N} f(I_q) \cdot f(I_{id})}{\sqrt{\sum_{i=1}^{N} f(I_q)^2} \sqrt{\sum_{i=1}^{N} f(I_{id})^2}}$$

f(Iq) represents fused feature vector of the query image and f(Iid) represents fused feature vector of individual images from the dataset, and N is the number of features. From the most similar image, the pose is extracted using Open pose. This pose is superimposed on the query image in the form of a skeleton.

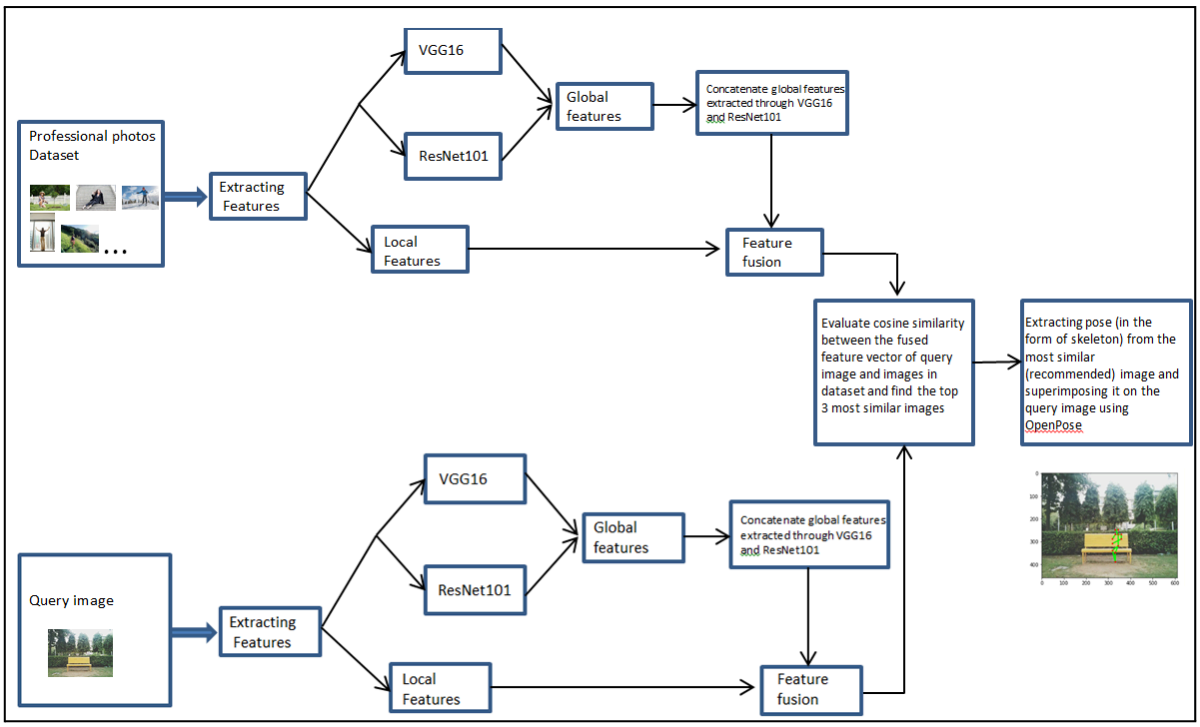
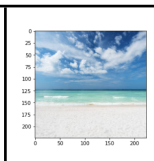
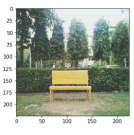
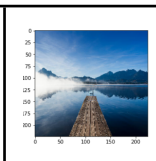
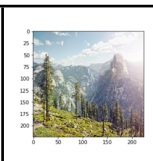
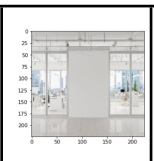
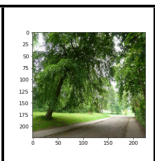
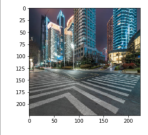
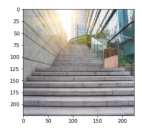
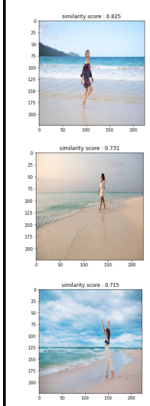
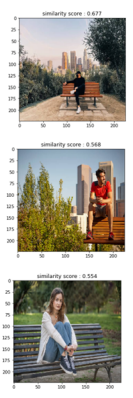
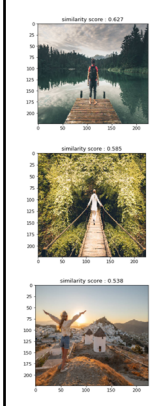
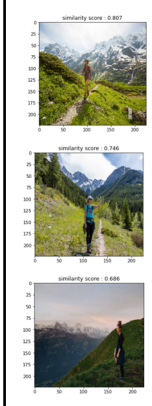
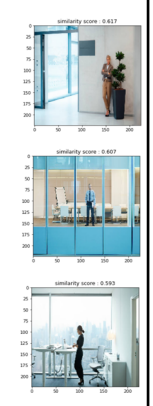
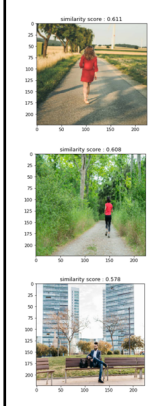
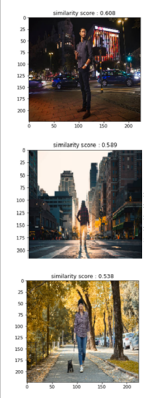


Figure 4: Framework of our approach

7. RESULTS

We tried to extract global features using various different backbone architectures like vgg16, vgg19, xception, resnet, inception_v3, inception_resnet_v2, mobilenet, densenet, nasnet, mobilenet_v2. Out of these models, VGG16, VGG19, ResNet101 and ResNet152 gave the best results. For the above four architectures for extracting global features, we obtained results using 3 different similarity metrics: 1) Cosine similarity; 2) Euclidean distance, and 3) Chi-Square distance.

The figures below showcase the results for VGG16, VGG19, ResNet101 and ResNet152 for cosine similarity, euclidean distance and chi squared similarity metrics in *Figure 5*, *Figure 6* and *Figure 7* respectively.

Query image -> Background architecture									
VGG16									

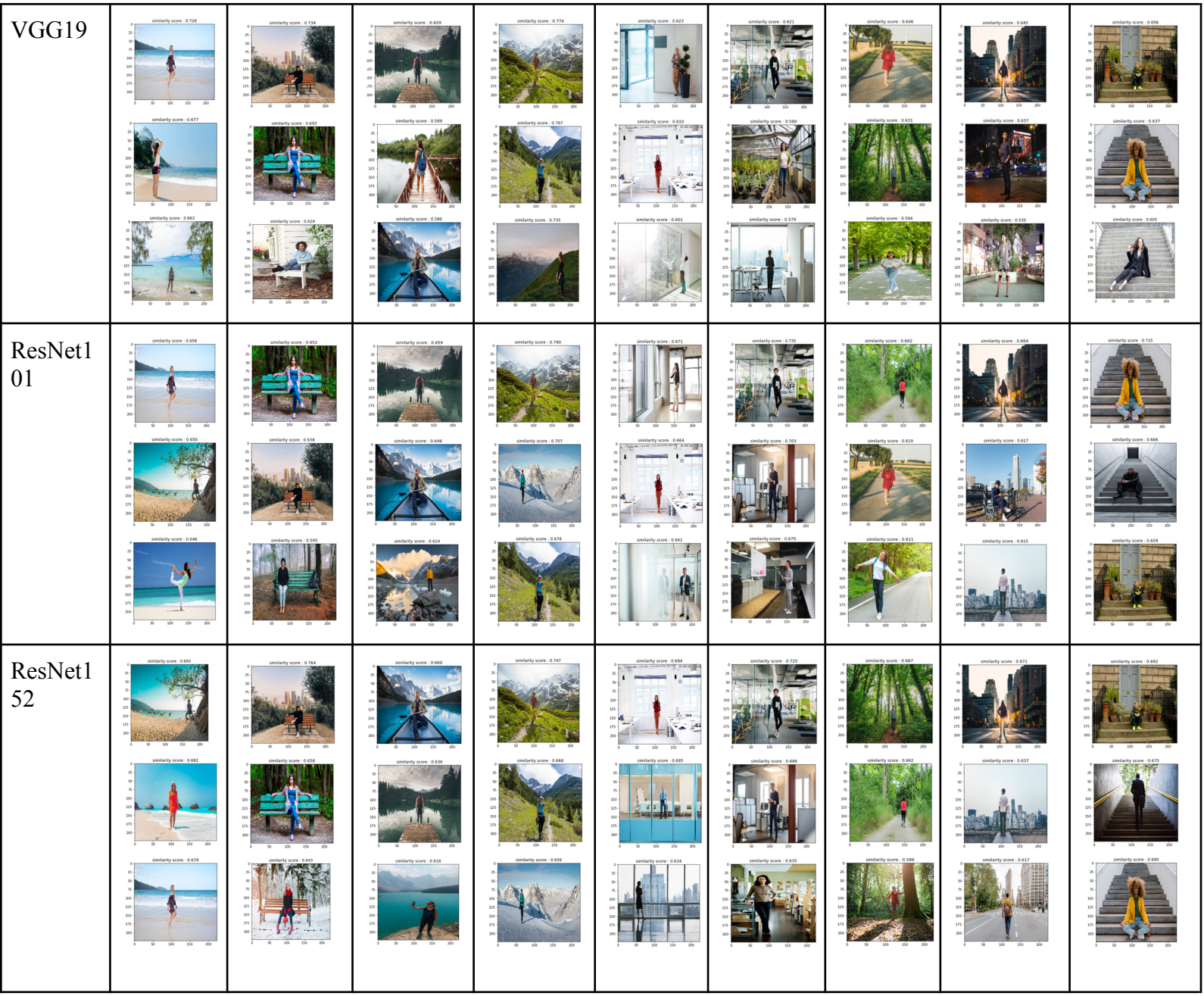
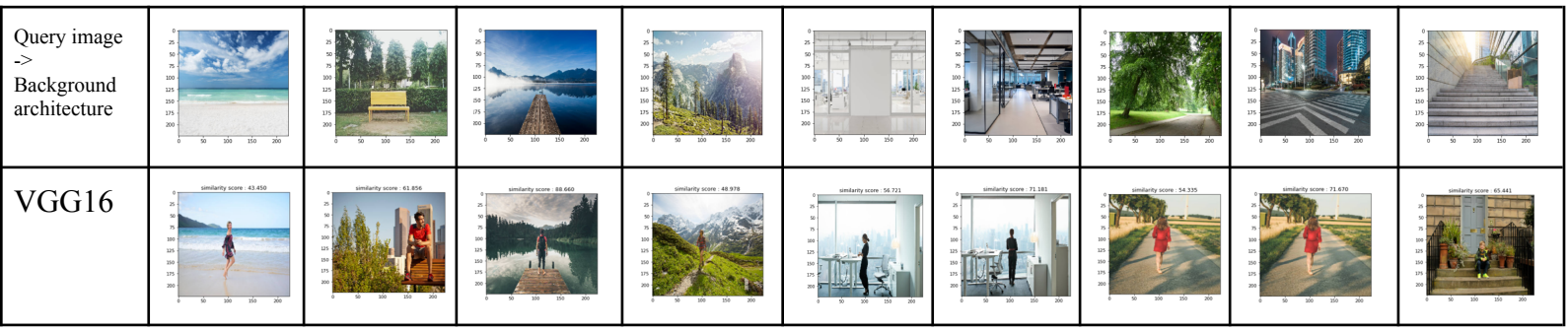


Figure 5: Top 3 recommended images for different CNN using cosine similarity index



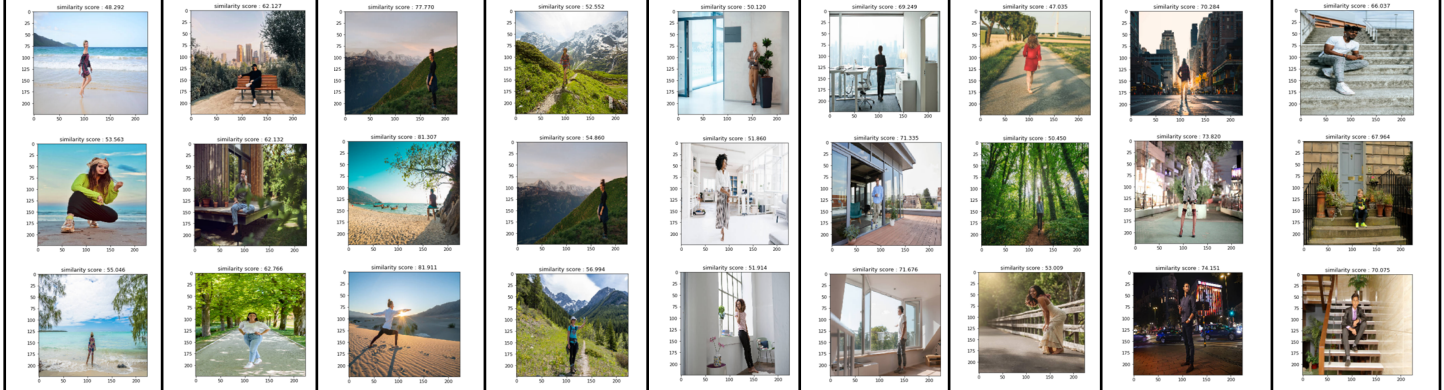
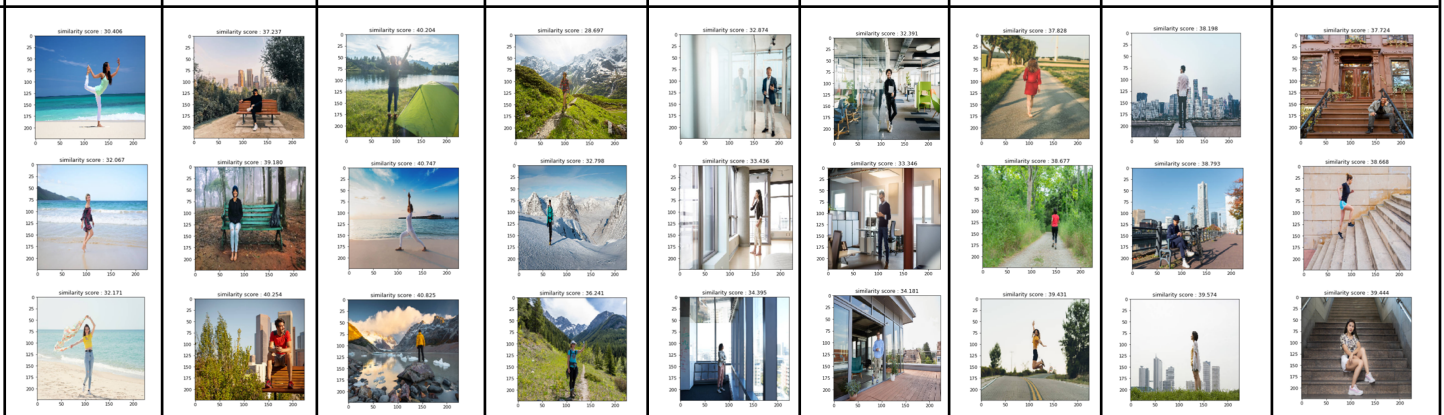
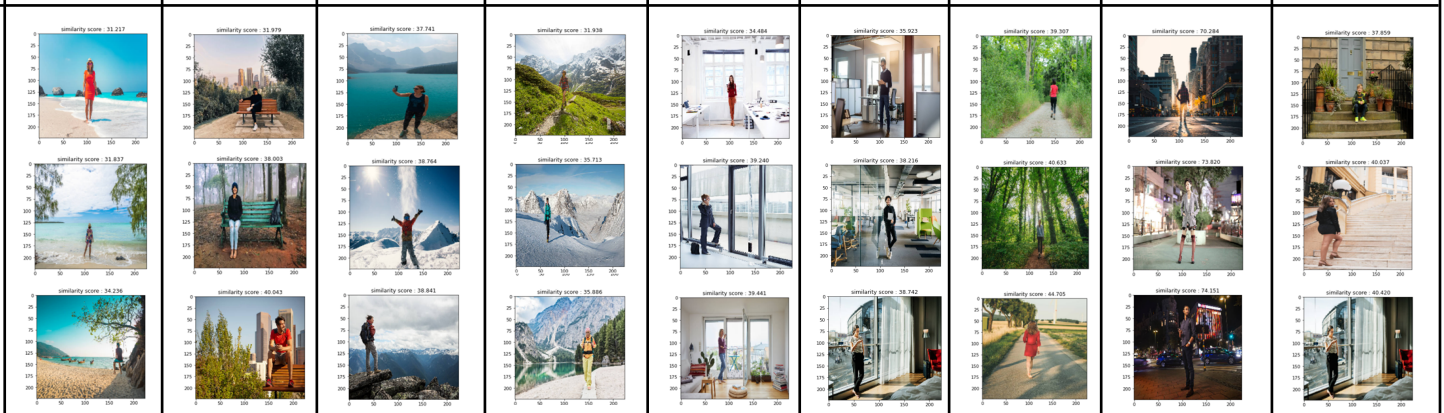
									
VGG19									
ResNet101									
ResNet152									

Figure 6: Top 3 recommended images for different CNN using Euclidean distance metric

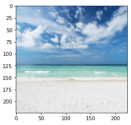
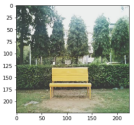
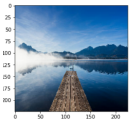
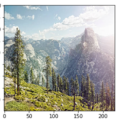
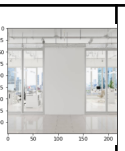
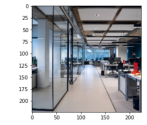
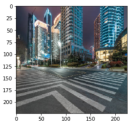
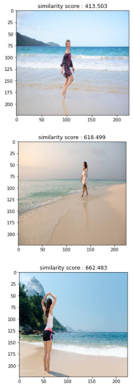
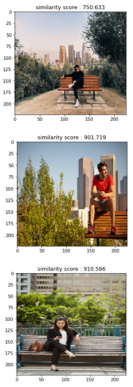
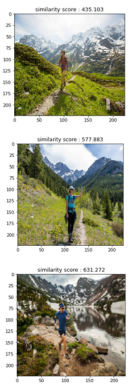
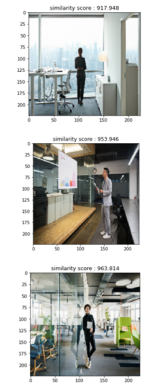
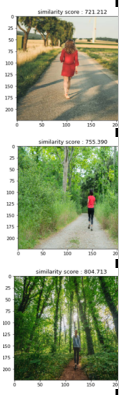
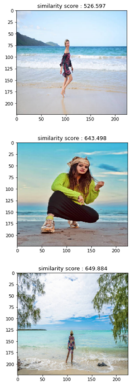
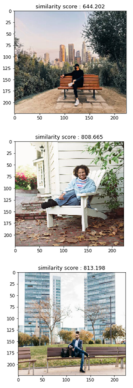
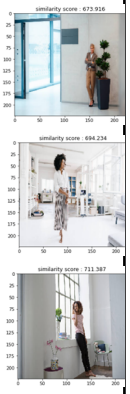
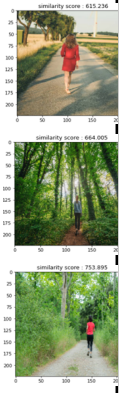
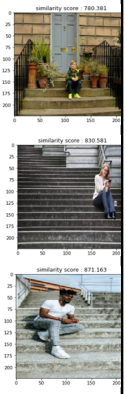
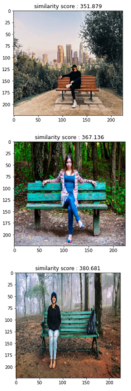
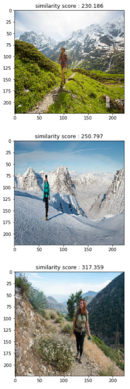
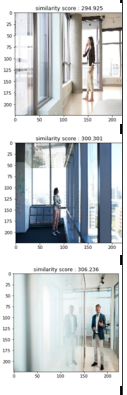
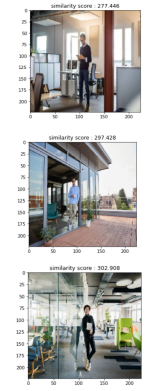
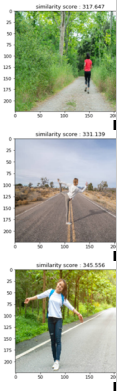
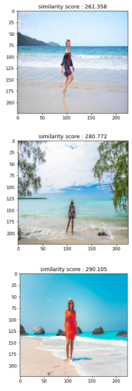
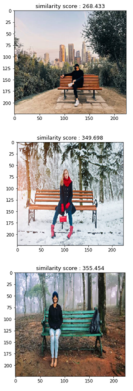
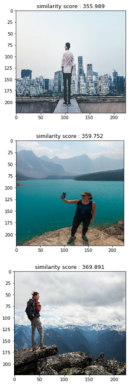
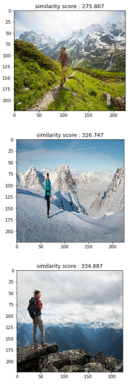
Query image -> Background architecture									
VGG16									
VGG19									
ResNet101									
ResNet152									

Figure 7: Top 3 recommended images for different CNN using Chi-Square distance metric.

After implementing the built algorithm on 4 different CNN architectures and 3 different similarity indexes, it was observed that the VGG16 model and ResNet101 model tend to perform better for almost all the test images as compared to the other architectures. Also it was observed that the cosine similarity metric gives a better similarity score. Hence, we tried to implement a hybrid model, which takes into picture the features extracted by both VGG16 and ResNet101 architecture as well as the local features, and uses cosine similarity for searching the most similar image to the query image. The results obtained for this hybrid model are quite satisfactory as shown in *Figure 8*.

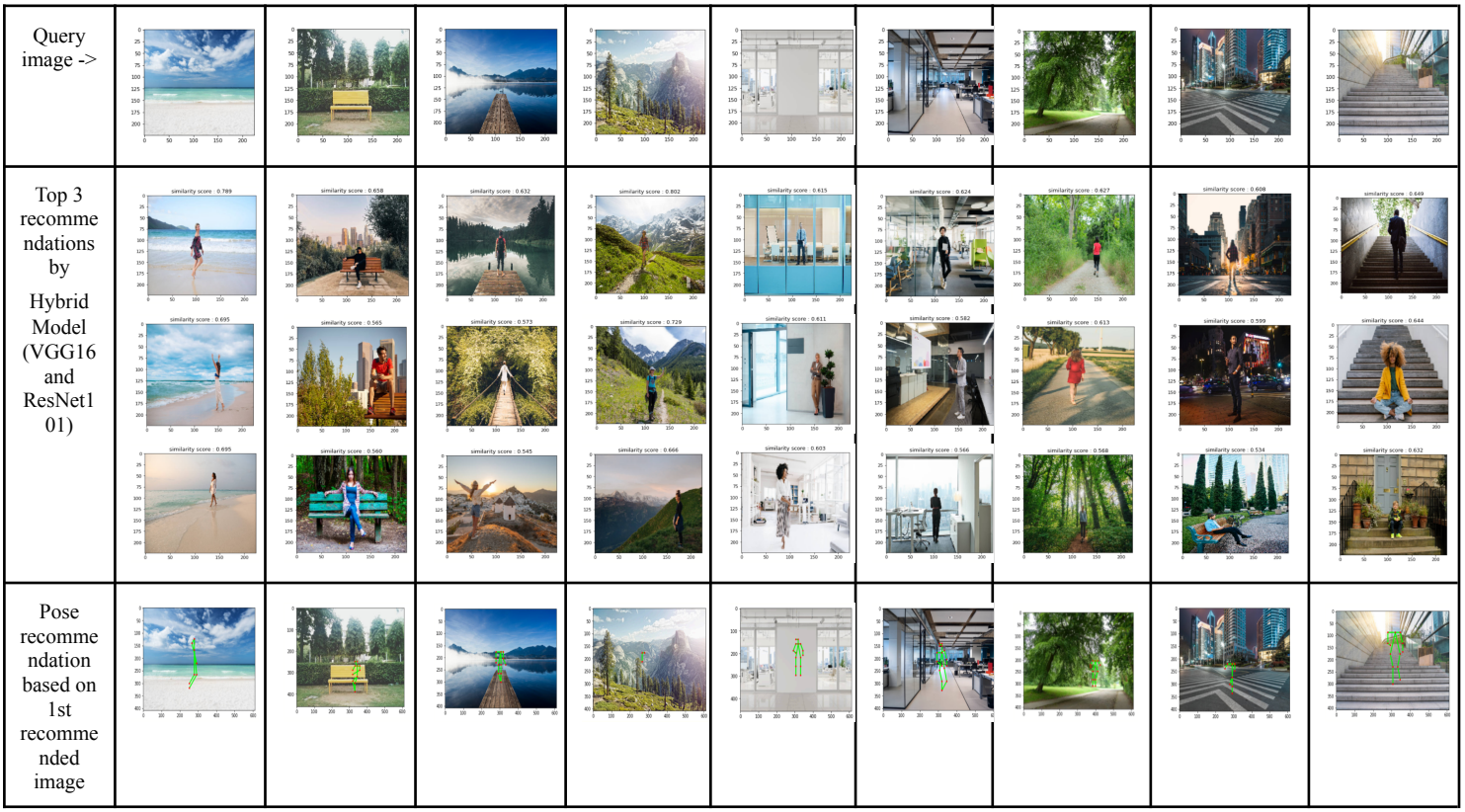


Figure 8: Top 3 recommended images for hybrid of VGG16 and ResNet 101 for global feature extraction and using Chi-Square distance metric.

User Study:

It is difficult to judge the artistry of photos and there is no qualitative standard which would evaluate algorithms used for recommending poses. In order to evaluate the performance of the proposed method, a subjective experiment was conducted. 25 participants (15 female and 10 male) from age group 18 to 22 years old were included in an experiment to judge the recommended photos. Each person rated the images according to the similarity of the recommended image to the query image. A total of 20 query images were judged for five levels of satisfaction.

Each measures the recommendation results of 20 images for four satisfaction levels. The four satisfaction levels are Very Satisfied, Satisfied, Slightly Satisfied, Not Satisfied. The statistical representation of user study is shown in *Figure 9*. The graph shows that the satisfaction rate (above satisfied) is 65.4%

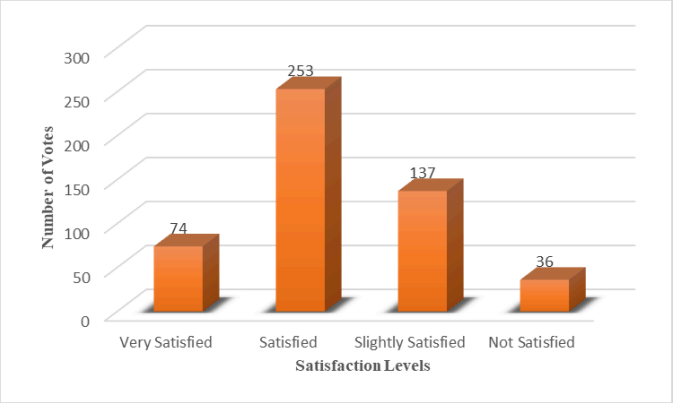


Figure 9: Results of the satisfaction levels of 25 users on 20 images. The results are summarised as Very Satisfied (14.8%), Satisfied(50.6%), Slightly Satisfied(27.4%), Not Satisfied(7.2%)

8. CONCLUSION AND FUTURE SCOPE

In this study, a novel approach for recommendation of poses for the input query image is proposed. A combination of global and local features is used for representing an image and recommending the most similar image. The feedback from users shows that the proposed algorithm gives satisfactory results.
The further scope of this approach includes recommending poses for real time query images uploaded from users' cameras. This approach is for single person pose recommendation, it can be further extended to multiple person pose recommendation. Also, a module for pose correction can be introduced which can act as a feedback for users after a user tries to imitate pose as recommended by this algorithm.

9. ACKNOWLEDGEMENT

We are grateful to our college and to the Department of Electronics and Telecommunication for providing us this opportunity to work on this project. It is their visionary objective to encourage students for this thesis-based project that has blossomed into this extraordinary opportunity.

To begin with, we sincerely thank our project guide Dr. Abhishek Bhatt for his guidance towards the design and development of the project. His constant motivation, support in bolstering requisite theoretical concepts along with invaluable tips and tricks proved to be crucial. We also would like to express our gratitude towards the technical and non-technical guidance provided by the faculties of our department.

The team is immensely grateful to the Head of the Department Dr. S. P. Mahajan. It is his support and encouragement to students to work on a project of their field of interest that has made this possible.

10. REFERENCES

[1] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu, “RMPE: Regional Multi-person Pose Estimation,” Nov. 2016, [Online]. Available: http://arxiv.org/abs/1612.00137.

[2] L. Pishchulin et al., “DeepCut: Joint subset partition and labeling for multi person pose estimation,” Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., vol. 2016-Decem, pp. 4929–4937, 2016, doi: 10.1109/CVPR.2016.533.

[3] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele, “DeeperCut: A Deeper, Stronger, and Faster Multi-person Pose Estimation Model BT - Computer Vision – ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI,” Arxiv, vol. 9910 LNCS, pp. 34–50, 2016, [Online]. Available: http://dx.doi.org/10.1007/978-3-319-46466-4_3.

[4] S.-E. Wei, V. Ramakrishna, T. Kanada, and Y. Sheikh, “Pose Machines :,” Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016.

[5] J. Carreira, P. Agrawal, K. Fragkiadaki, and J. Malik, “Human pose estimation with iterative error feedback,” Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., vol. 2016- Decem, pp. 4733–4742, 2016, doi: 10.1109/CVPR.2016.512.

[6] Newell, K. Yang, and J. Deng, "Stacked Hourglass Networks for Human Pose Estimation," in ECCV, 2016.

[7] Z. Cao, G. Hidalgo, T. Simon, S. E. Wei, and Y. Sheikh, “OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 43, no. 1, pp. 172–186, 2021, doi: 10.1109/TPAMI.2019.2929257.

[8] K. Sun, B. Xiao, D. Liu, and J. Wang, “HRNet1,” Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., vol. 2019-June, pp. 5686–5696, 2019.

[9] F. Zhang, X. Zhu, H. Dai, M. Ye, and C. Zhu, “Distribution-Aware Coordinate Representation for Human Pose Estimation,” Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., pp. 7091–7100, 2020, doi: 10.1109/CVPR42600.2020.00712.

[10] Tong Hao, QianWang, DanWu, JinSheng Sun, “Adaptive recommendation for photo pose via deep learning “

[11] Farshid Farhat, Mohammad Mahdi Kamani, Sahil Mishra, James Z. Wang, “Intelligent Portrait Composition Assistance - Integrating Deep-learned Models and Photography Idea Retrieval”

[12] Yanhao Zhang, Xiaoshuai Sun, Hongxun Yao, Lei Qin, Qingming Huang, ”Aesthetic Composition Representation For Portrait Photographing “

[13]https://pyimagesearch.com/2014/12/01/complete-guide-building-image-search-engine-python-opencv/

[14] https://www.gettyimages.in/

[15] Bin Cheng, Bingbing Ni, ShuiCheng Yan and Qi Tian. Learning to Photograph. ACM MM 2010.

[16] T. L. Munea, Y. Z. Jembre, H. T. Weldegebriel, L. Chen, C. Huang, and C. Yang, “The Progress of Human Pose Estimation: A Survey and Taxonomy of Models Applied in 2D Human Pose Estimation,” IEEE Access, vol. 8, pp. 133330–133348, 2020, doi: 10.1109/ACCESS.2020.3010248.
[17] Q. Dang, J. Yin, B. Wang, and W. Zheng, “Deep learning based 2D human pose estimation: A survey,” Tsinghua Sci. Technol., vol. 24, no. 6, pp. 663–676, 2019, doi: 10.26599/TST.2018.9010100.

[18] S. C. Babu, "A 2019 guide to Human Pose Estimation with Deep Learning," 2019. [Online]. Available: https://nanonets.com/blog/human-pose-estimation-2d-guide/

[19] D. Mwiti, "A 2019 Guide to Human Pose Estimation," 2019. [Online]. Available: https://heartbeat.fritz.ai/a2019-guide-tohuman-pose-estimation-c10b79b64b73

[20] M. Patel and N. Kalani, “A survey on Pose Estimation using Deep Convolutional Neural Networks,” IOP Conf. Ser. Mater. Sci. Eng., vol. 1042, no. 1, p. 012008, 2021, doi: 10.1088/1757-899x/1042/1/012008.

[21] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "MPII Human Pose Dataset." [Online]. Available: http://humanpose.mpi-inf.mpg.de/

[22] T. Y. Lin et al., “Microsoft COCO: Common objects in context,” Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 8693 LNCS, no. PART 5, pp. 740–755, 2014, doi: 10.1007/978-3-319-10602-1_48.

[23] E. M., V.-G. L., W. C. K. I., W. J., and Z. A., “The Pascal Visual Object Classes (VOC) Challenge,” Int. J. Comput. Vis., vol. 88, no. 2, pp. 303–338, 2010.

[24] G. Varol et al., "SURREAL: Learning from Synthetic Humans," in CVPR, 2017. [Online]. Available: https://www.di.ens.fr/willow/research/surreal/data/

[25] L. Sigal, A. O. Balan, and M. J. Black, “HumanEva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion,” Int. J. Comput. Vis., vol. 87, no. 1–2, pp. 4–27, 2010, doi: 10.1007/s11263-009-0273-6.

[26] K. Sun, B. Xiao, D. Liu, and J. Wang, “HRNet1,” Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., vol. 2019-June, pp. 5686–5696, 2019.

[27] F. Zhang, X. Zhu, H. Dai, M. Ye, and C. Zhu, “Distribution-Aware Coordinate Representation for Human Pose Estimation,” Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., pp. 7091–7100, 2020, doi: 10.1109/CVPR42600.2020.00712