Pose Estimation Keypoints in Age Recognition of Full Body Image

Rachad Lakis*, Joseph Constantin †‡ *, Ibtissam Constantin† *, Vinh Truong Hoang‡, Yassine Ruichek §
* ESIB, Saint Joseph University Beyrouth, Lebanon
† LaRRIS Laboratory, Lebanese University, Fanar, Lebanon
‡ Ho Chi Minh City Open University, Ho Chi Minh City, Viêt Nam
§ CIAD Laboratory, UTBM University Montbéliard, France

***Abstract*—Age recognition based on an image has many potential applications, such as developing intelligent human-machine interfaces and improving safety in various fields, such as transportation and security. Most age classification learning models use a person's face to estimate the age category. However, images are usually captured in-the-wild, where often no near-frontal information is available. Also, images are taken under different illumination conditions and different camera viewing angles, providing poor visual quality. To address this problem, we proposed a novel deep learning algorithm for age recognition of adults and children using images in which faces are not well recognized. Our results clearly show that the addition of Pose Estimation Keypoints leads to significant improvement in age classification accuracy, precision, recall, and F1-score using deep learning models with simple architecture for online decision making . It can be seen that Pose Estimation Keypoints can have important implications for a range of applications in areas such as CCTV footage, recommendation of videos and advertisements according to the target audience, granting additional privileges to individuals depending on their age and even on social media where there are age restrictions for viewing or using certain content on the social platforms.**

***Index Terms*—Age Estimation, Deep Learning, Pose Estimation Keypoints.**

I. Introduction

Predicting the age of individuals from images is an important task in many applications such as advertising [1], and surveillance [2]. Many conventional approaches to age classification have focused on two phases: The first is extracting image features and representations for age, and the second is learning an age estimator based on the image features. Various age prediction models have been developed. The anthropometric models, which use the distance between viewpoints to understand the geometric structure of humans, can help distinguish between different ages [3]. Unlike anthropometric models, texture-based models can perform well on images acquired under uncontrolled conditions, but they are unable to distinguish different shapes and distances between facial features [4]. Active appearance models use a dimensionality reduction algorithm to learn the extracted texture and shape features [5]. However, the dimensionality reduction of the features causes wrinkles to go unnoticed. The aging pattern subspace has been proposed to identify a person's aging pattern based on a set of facial images, but does not work well on wrinkles. Age Manifold focuses on treating aging patterns as a trend for multiple individuals at different ages, rather than finding a specific aging pattern for each person. However, these techniques project high-dimensional data into a unit hypersphere. Unsupervised dimensionality reduction techniques are not effective for dealing with distinguishing features [6]. Recent advances in computer vision have allowed researchers to consider other parts of the body. Whole body images contain information about a person's age, including posture, body shape, and other visual cues that can provide valuable insights into identifying a person's age. Since the deep convolution neural network has demonstrated its ability and intelligence in many feature extraction and segmentation applications, various techniques have been developed for facial feature extraction by tuning a convolution neural network from scratch [7]. However, tuning a network from scratch is computationally expensive. Pre-built models are used to fine-tune a network that has been trained for a more complicated task of extracting features from an image [8]. These models have shown excellent accuracy in face recognition. However, these models are based only on facial images, and the available public datasets are not suitable for the case of whole-body images. It becomes difficult for the models to accurately predict a person's age when unclear or blurred images are given as input and especially when images are captured in-the-wild, where often no near-frontal information is available. In this work, our contribution is to recognize the age of a person even if his face is not well recognized. This problem is encountered where images are taken under different illumination conditions and different camera viewing angles, providing poor visual quality. To address this challenge, transfer learning-based approaches have been proposed to adapt already trained models to new domains. We seek to explore the potential benefits of incorporating Pose Estimation Keypoints into age prediction of body images [9]. Pose Estimation Keypoints is a computer vision technique that detects specific points or landmarks on an object or a person's body. To test our algorithm, we fine-tuned several Deep Neural Network models on a dataset of full-body images provided by AI Directions, an artificial intelligence research and consulting agency, with and without the addition of Pose Estimation Keypoints. We then compared the performance of these models using various evaluation metrics such as accuracy, precision, recall and F1-score. Next, we performed a comparative study between the state of the art of existing

models and other Convolution Neural Networks with simple architecture tuned from scratch. Our study highlights the potential of Pose Estimation Keypoints as an important tool for improving the accuracy, precision, recall and F1-score of age classification for full-body images in case of models with simple architecture used for online decision making. These findings are relevant to a range of applications including video recommendation, advertising, and even social media.

This paper is organized as follows: Section 2 presents the methodology used to create the dataset, the Pose Estimation Keypoints algorithm, the implementation of the models and, the evaluation metrics used in the paper. Section 3 presents the experimental study and the results obtained. Finally, section 4 presents the conclusions and future perspectives.

II. Methodology

A. Dataset

Obtaining a suitable training dataset is the most important step in building a machine learning model. Various datasets have been used for age estimation. However, these datasets were originally introduced for face recognition tasks, and thus are not suitable for our case. The Inria person dataset is the only dataset that contains whole body images but it consists of only 800 images and is used in the literature to evaluate pedestrian recognition algorithms [10]. The dataset used in this study consists of adult and child images and was provided by AI Directions agency (Figure 1). However, this dataset was unbalanced, containing 95% of adult images, and only 5% of child images. The model cleverly labeled all images with adult labels. While this quickly leads to high accuracy, it is a false victory. After implementing and testing many models, the accuracy of all models was more than 95% on the learning set, but this was because more than 90% of the dataset belonged to one class, which became apparent when we tested the Accuracy, Precision, Recall, and F1-score of the models. A good solution was to expand the dataset and give a higher weight to the class with the lower occurrence. To deal with the unbalance of classes, we used the class weighting technique. When the classes are not balanced, the classifier should heavily weight the few available examples. This was achieved by passing Keras weights for each class

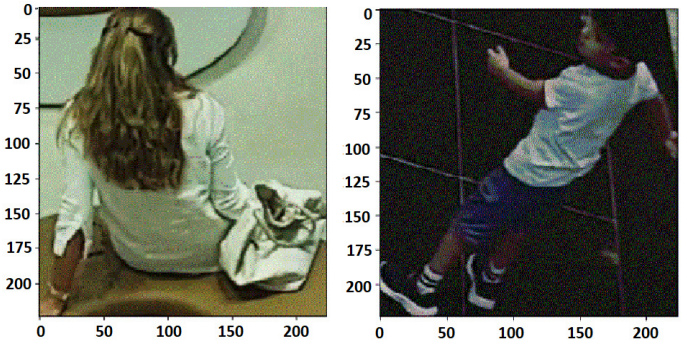

Fig. 1. Artificial intelligence research and consulting agency AI Directions Images.

via parameters. This caused the model to pay more attention to examples from an underrepresented class. We also found that the first dataset did not provide enough information to to evaluate the performance of the learning model on real-world data. Therefore, we created a new dataset to address the limitation we observed in the first set. This limitation stemmed from the fact that a significant portion of the images were primarily focused on specific clothing traditions. In order to expand the scope and diversity of our dataset, we set out to find additional visual resources online. We created a new dataset by searching for new real-world images showing the whole body of a person or child in different positions. We cleaned all images and resized them to (224,224,3). The new dataset consisted of 5349 images for adults, and 6320 images for children, divided into 80% for training, 10% for validation and 10% for testing.

B. Pose Estimation Keypoints

Pose Estimation Keypoints is the task that employs computer vision techniques to estimate the configuration of the human body in a given image or a sequence of images. It has the advantage over object detection models, which can locate objects in an image but provide only coarse-grained localization. We can categorize the current approaches to Pose Estimation Keypoints into bottom-up and top-down. In the bottom-up methods, body Keypoints are first detected in an image, without knowing the number or location of person instances or to which person instances these Keypoints belong. Next, the detected Keypoints are grouped and assigned to person instances [11]. Recent works densely regress a set of pose candidates, where each candidate consists of the Keypoint positions that might be from the same person [12]. Unfortunately, the regression quality is not high and a post-processing scheme is usually adopted to improve the regression results.

Top-down methods first detect person instances in the input image [13]. They usually use a standard object detector to obtain person boxes. Next, top-down methods estimate a single person pose for each person box cut out. Since a pose estimation model is run for each person, top-down methods tend to be slow on average. They become much slower as the number of persons in an image increases. Other method include developing a multi-task learning architecture combining detection and segmentation to improve Keypoint localization [14]. In this paper, we used a bottom up approach in which all Keypoints for each person in the image are predicted by fully convolution. It predicts the relative displacement between each pair of Keypoints and also proposes a novel recurrent scheme that significantly improves the accuracy of the predictions for large distance. Once the Keypoints are located, they are grouped into instances using a greedy decoding process. This approach starts from the most confident detection, as opposed to always starting from a distinguished landmark such as the nose, so it works well when clutter is present [15]. We used the Movenet Thunder, a Deep Learning-based model that estimates peoples postures from video and image data [16]. This model can run at over 200 frames per second

on a single CPU core and uses a multi-stage architecture consisting of multiple lightweight neural networks. The first stage of Movenet Thunder is a person detector that identifies the presence of human bodies in the input image or video. The second stage is a body pose regression that estimates the 3D poses of the detected people [17]. In our case, the cameras capture whole body images, and sometimes the faces are not well recognized. Our learning model tries to exploit the fact that the ratio between the body parts of children and adults is different to make an image prediction of age. We converted the image into a tensor and passed it to the model for Pose Estimation Keypoints. We got the resulting Keypoints, and then added a circle for each Keypoint and a line for each connection of the Keypoints. The steps to compute the Pose Estimation Keypoints are explained in algorithm 1.

C. Deep Learning models

Convolution neural network (CNN) is a feed-forward neural network capable of extracting features from data with convolution structures [18]. The goal is to create an accurate model that is not too large. A large model can take a lot of time in inference mode, and this is not compatible with the online use case. Therefore, we first chose the $VGG19$ network defined in [19], which uses an architecture with very small convolution filters and yields significant improvements over previous configurations by increasing the depth of the weighting layers to 19. Next, we chose the $ResNet50$ model [8] where the layers were reformulated to learn residual functions referenced to the layer inputs, rather than learning unreferenced functions. It turns out that these residual networks are easier to optimize, and can gain accuracy by being significantly deeper. We also found that $MobileNetV2$ [20] is a great candidate because it has a small size and can produce great results. Its architecture is based on an inverted residual structure, where the shortcut connections are between the thin bottleneck layers. This model has attracted attention due to its unique design, that provides high accuracy with lower computational complexity compared to models such as $VGG19$ [21].

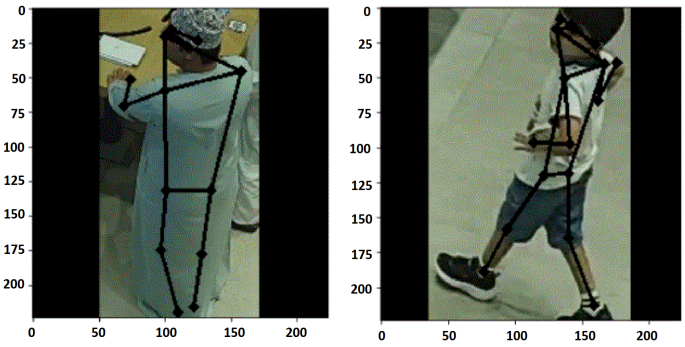

Fig. 2. Pose Estimation Keypoints in Age Estimation.

Algorithm 1 *Algorithm for computing the Pose Estimation Keypoints*

Require: *A dataset containing images of different adults and children.*

for all images **do**
  $Image \leftarrow Convert\_Tensor(Image)$
  $Image \leftarrow Resize\,(image, (256, 256))$
  $Keypoints \leftarrow passing\_pose\_model(image)$
  for all points in Keypoints **do**
    Draw_circle(point)
  end for
  for all edge in edges_between_Keypoints **do**
    Draw_line(edge)
  end for
end for
Return the Pose Estimation Keypoints.

D. Evaluation Metrics

The evaluation metric plays a crucial role in finding the optimal classifier during classification training [22]. Since the dataset is not balanced, the accuracy of the model alone is not sufficient. In this study, four evaluation indices were selected to assess performance, including accuracy, precision, recall, and F1-Score. Precision measures the proportion of positively predicted labels that are actually correct. It represents the ratio of true positive to the sum of true positive and false positive. In contrast, recall represents the model's ability to correctly predict the positives out of actual positives. It represents the ratio of true positive to the sum of true positive and false negative. F1-Score, also known as the balanced score, is defined as the harmonic mean of the precision and recall rates. Accuracy is a machine learning classification model performance metric that is defined as the ratio of true positives and true negatives to all positive and negative observations. It represents the ratio of the sum of true positive and true negatives out of all the predictions. These evaluation metrics were calculated between the predicted labels and the ground truth. The calculation equations for the indicators are shown in Table I .

TABLE I
EVALUATION METRICS. TP, FP, FN, AND TN ARE THE TRUE POSITIVE, FALSE POSITIVE, FALSE NEGATIVE AND TRUE NEGATIVE CLASSIFICATION.

Evaluation Criteria	Formula
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$
Precision	$\frac{TP}{TP+FP}$
Recall	$\frac{TP}{TP+FN}$
F1-score	$2 \times \frac{Precision \times Recall}{Precision+Recall}$

III. EXPERIMENTAL RESULTS

The code was run using the GPU T4 x2 Kaggle environment due to the computational intensity of the deep learning algorithm used in image age classification, transfer learning, and Pose Estimation Keypoints. In the classification task, we tested different Deep Learning models with different configurations, including different number of trainable layers and Pose Estimation Keypoints to compare their performance

and determine the best approach for each task. In our work, we conducted three experiments:

In the first experiment, we performed for the VGG19 [19], Resnet50 [8] and MobileNetV2 [20] Convolution Neural Networks transfer learning and fine tuning. The base model trained in the ImageNet database was loaded without the dense layers. Next, we froze the base model and added a Global Average Pooling 2D layer, a dropout layer with probability 0.2, a fully connected layer with 64 neurons using Relu activation function, a dropout layer with probability 0.5 and a fully connected layers with one output neuron using sigmoid activation function (figure 3). The fully connected layers

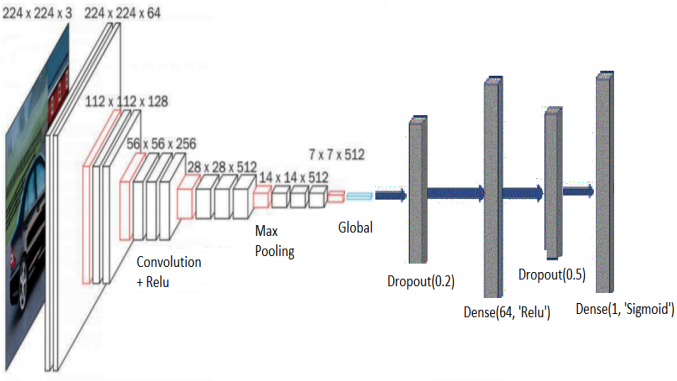

Fig. 3. The convolution base of the VGG19 model used for transfer learning and fine tuning.

were trained with the Adam optimizer. Then 10% of the base model was unfrozen and trained again for 20 epochs. Table II shows the accuracy, precision, loss, F1-score and the time consumption on a block of 100 images for the CNNs. It is clear that the VGG19 model using the Pose Estimation Keypoints is slightly better than the other Convolution Neural networks. In some cases, we cannot see a big improvement, because the CNN is already having very high metrics, and any added improvement could not be valid since the model achieves a higher performance. We also plot the variation of Loss, Accuracy, Precision and Recall during learning for the VGG19 model with Pose Estimation Keypoints that achieves the best performance (Figure 4). The F1-score is computed based on the values of Precision and Recall.

In the second experiment, we unfroze all layers of the models and then trained the models for 8 epochs using the Adam optimizer. Table III shows the results of loss, accuracy, precision, Recall, and F1-score for the models. It can be seen that the VGG19 model with Pose Estimation Keypoints gives similar results than the other Convolution Neural Network models. Moreover, it is shown that the performance of the model with Pose Estimation Keypoints is achieved with a non-significant addition of time cost equal to 1second on a block of 100 images. We also plot the variation of Loss, Accuracy, Precision and Recall during learning for the VGG19 model with Pose Estimation Keypoints (Figure 5). This experiment did not show the effectiveness of the Pose Estimation Keypoints algorithm because all these models are very complex and

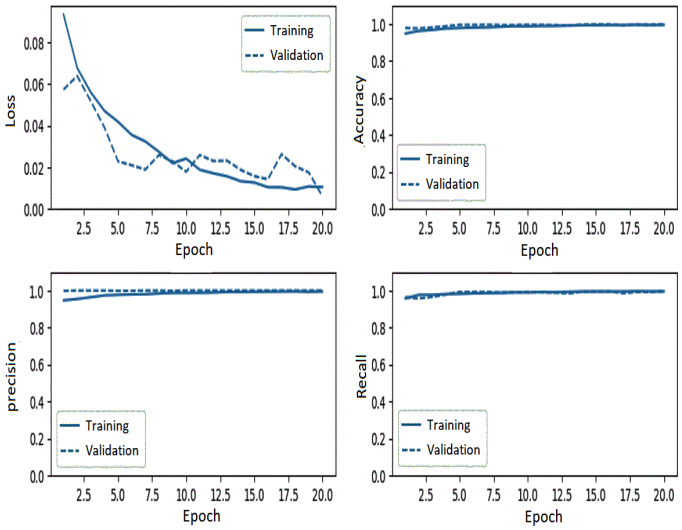

Fig. 4. Variation of Loss, Accuracy, Precision and Recall during learning for the VGG19 model using Pose Estimation Keypoints (First experiment).

effective. So, we tried to build simple Convolution Neural Network to show the effectiveness of the new algorithm and to minimize the complexity of the CNN.

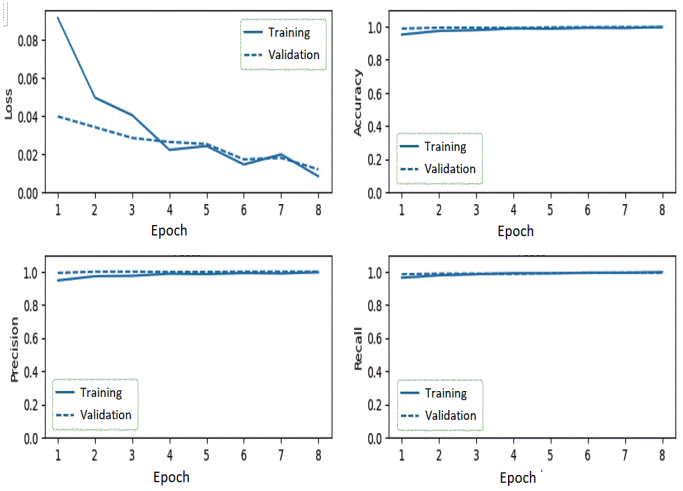

Fig. 5. Variation of Loss, Accuracy, Precision and Recall during learning for the VGG19 model using Pose Estimation Keypoints (Second experiment).

Finally, to show the advantage of the Pose Estimation Keypoints, we performed a comparative study between the state of the art of existing CNNs and other CNNs with simple architecture tuned from scratch. The architectures of the four developed CNNs are given in Table III. We trained the Convolution Neural Networks for a maximum of 40 epochs with the Adam optimizer, but we stopped the process if there was no improvement during the training. We configured a stop after 6 epochs if there was no improvement in validation loss. Table V shows the accuracy, precision, loss, F1-score and the time consumption on a block of 100 images for the CNNs. We show that CNN 2 using pose estimation Keypoints as input

TABLE II
FIRST EXPERIMENT RESULTS FOR ALL MODELS WITH AND WITHOUT SINGLE POSE ESTIMATION KEYPOINTS

Model	Number of Parameters	Pose Keypoints	Loss	Accuracy	Precision	Recall	F1-score	Time(s)
VGG19	**20057281**	No	0.0140	0.99	1.0	0.99	0.99	16.31
		Yes	**0.0013**	**1.0**	**1.0**	**1.0**	**1.0**	**17.31**
ResNet50	23718913	No	0.0035	0.99	1.0	0.99	0.99	6.05
		Yes	0.0090	0.99	1.0	0.99	0.99	6.5
MobileNetV2	2340033	No	0.0899	0.97	1.0	0.95	0.97	1.91
		Yes	0.0294	0.98	1.0	0.98	0.98	2.16

TABLE III
SECOND EXPERIMENT RESULTS FOR ALL MODELS WITH AND WITHOUT POSE ESTIMATION KEYPOINTS

CNN	Pose Keypoints	Loss	Accuracy	Precision	Recall	F1-score
VGG19	**No**	**0.0001**	**1.0**	**1.0**	**1.0**	**1.0**
	Yes	**0.0001**	**1.0**	**1.0**	**1.0**	**1.0**
ResNet50	No	0.0027	0.99	1.0	0.99	0.99
	Yes	0.0046	0.98	1.0	0.98	0.98
MobileNetV2	No	0.0579	0.98	1.0	0.97	0.98
	Yes	0.0292	0.99	1.0	0.99	0.99

TABLE IV
ARCHITECTURE OF THE MODELS TRAINED FROM SCRATCH.

Layers	CNN 1	**CNN2 2**	CNN 3	CNN 4
L1	Input (224,224,3)	**Input (224,224,3)**	Input (224,224,3)	Input (224,224,3)
L2	Conv. (224,224,32)	**Conv. (224,224,32)**	Conv. (224,224,32)	Conv. (224,224,32)
L3	Norm. (224,224,32)	**Norm. (224,224,32)**	Norm. (224,224,32)	Norm. (224,224,32)
L4	Conv. (224,224,32)	**Conv. (224,224,32)**	Conv. (224,224,32)	Conv. (224,224,32)
L5	Norm. (224,224,32)	**Norm. (224,224,32)**	Norm. (224,224,32)	Norm (224,224,32)
L6	M. Pooling (112,112,32)	**M. Pooling (112,112,32)**	M. Pooling (112,112,32)	M. Pooling (112,112,32)
L7	Conv. (112,112,32)	**Conv. (112,112,32)**	Conv. (112,112,32)	Conv. (112,112,32)
L8	Norm. (112,112,32)	**Norm. (112,112,32)**	Norm. (112,112,32)	Norm. (112,112,32)
L9	Av. Pooling (32)	**Conv. (112,112,32)**	Conv. (112,112,32)	Conv. (112,112,32)
L10	Dense (32)	**Norm. (112,112,32)**	Norm. (112,112,32)	Norm. (112,112,32)
L11	Dropout (32)	**Av. Pooling (112,112,32)**	M. Pooling (56,56,32)	M. Pooling (56,56,32)
L12	Dense (1)	**Dense (32)**	Conv. (56,56,32)	Conv. (56,56,32)
L13	None -	**Dropout (32)**	Norm. (56,56,32)	Norm. (56,56,32)
L14	None -	**Dense (1)**	Av. Pooling (32)	Conv. (56,56,32)
L15	None -	**None -**	Dense (32)	Norm. (56,56,32)
L16	None -	**None -**	Dropout (32)	AV. Pooling (32)
L17	None -	**None -**	Dense (1)	Dense (32)
L18	None -	**None -**	None -	Dropout (32)
L19	None -	**None -**	None -	Dense (1)

is better in convergence than the other Convolution Neural Networks. Moreover, it is clear that the performance of the model using pose estimation Keypoints is achieved with a non-significant addition of time cost equal to 0.3 second on a block of 100 images. We plot also the variation of Loss, Accuracy, Precision and Recall for CNN 2 that assures the best performance (Figure 6). It is shown that the training metrics increase linearly over time, whereas the validation metrics oscillates because we have a few training samples in the learning base which leads to overfitting. It is shown that

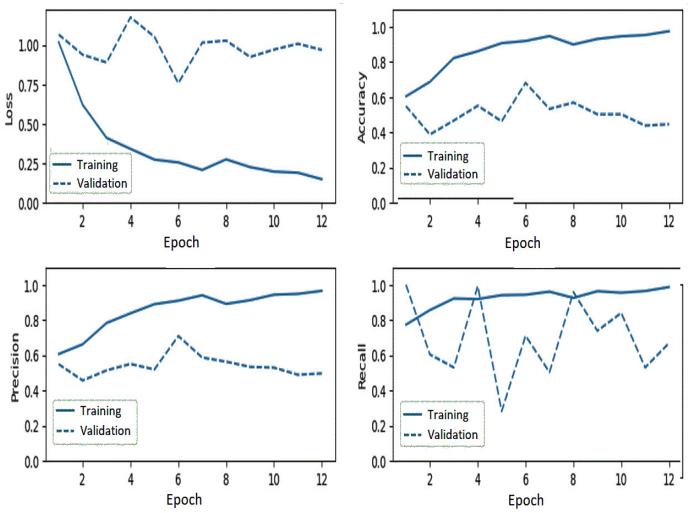

Fig. 6. Variation of Loss, Accuracy, Precision and Recall during learning for CNN 2 using Pose Estimation Keypoints (Third experiment).

a simple CNN used for online decision making shows a great improvement in the metrics after adding the Pose Estimation Keypoints and the more the CNN becomes complex, the more the difference between CNN before and after adding the Pose Estimation Keypoint becomes nearly null. In fact, a complex CNN does not need the Pose Estimation Keypoints. Its complexity is more than enough to attain the best metrics without the use of it, but for a simple CNN, it shows a very important improvement.

TABLE V
THIRD EXPERIMENT RESULTS FOR ALL CNNS WITH AND WITHOUT POSE ESTIMATION KEYPOINTS

Model	Number of Parameters	Conv.	Layers	KeyPoints	Loss	Accuracy	Precision	Recall	F1 score	Time(s)
CNN 1	20865	3	12	No	1.4474	0.43	0.54	0.08	0.13	1.50
		3	12	Yes	0.9474	0.45	0.54	0.34	0.41	1.82
CNN 2	**30241**	4	14	No	0.9553	0.44	0.51	0.66	0.57	1.69
		4	**14**	**Yes**	**0.6753**	**0.73**	**0.70**	**0.97**	**0.79**	**1.99**
CNN 3	39617	5	17	No	0.7831	0.64	0.64	0.85	0.73	1.70
		5	17	Yes	0.7361	0.64	0.61	0.97	0.75	2.17
CNN 4	48993	6	19	No	0.9001	0.64	0.62	0.97	0.75	1.81
		6	19	Yes	1.2113	0.64	0.62	0.97	0.76	2.23

IV. CONCLUSION

The main idea of this paper is to develop a Convolution Neural Network for detecting age categories on whole-body images where faces are not well recognized, and to investigate the potential benefits of including Pose Estimation Keypoints in this task. We successfully achieved our goal by building a highly accurate Deep Learning CNN. Through extensive research, we were able to gain valuable insight into the impact of integrating these Keypoints as input to the CNNs. Our results show that the integration of Pose Estimation Keypoints led to significant improvements in age classification accuracy, precision, recall, and F1-score in the case of a Convolution Neural Network with simple architecture used to make online decisions on images scrolling on a video system. Future work will require evaluation of the CNN using a full frame with a large number of images without faces. Moreover, it may include comparing the performance of the obtained model with other pre-trained deep learning models under testing on a huge number of real images. In addition, an extension of this work may be to consider more classes, where each class corresponds to an age range, this may benefit some applications that requires a knowledge on specific ranges of age.

ACKNOWLEDGMENT

This project has been funded with support from the HCMC Open University. We would like to thank the Artificial Intelligence Research and Consulting Agency AI Directions for providing us with the data used in our experiments.

REFERENCES

[1] Karen A. , Christopher M. , Aherman V. , Qomariyah N. and Astriani M. , Analyzing the Impact of Age and Gender for Targeted Advertisements Prediction Model, International Conference on Data Science and Its Applications (ICoDSA), Bandung, Indonesia, 2022, pp. 70-75, doi: 10.1109/ICoDSA55874.2022.9862531.

[2] Vasavi S., Vineela P. , Venkat Raman S., Age Detection in a Surveillance Video Using Deep Learning Technique. *SN Computer Science* 2021, *2*, , 4, https://doi.org/10.1007/s42979-021-00620-w.

[3] Fu, Y.; Hospedales, T.M.; Xiang, T.; Gong, S.; Yao, Y. Interestingness Prediction by Robust Learning to Rank. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014, pp 488–503.

[4] Unnikrishnan, A.; Ajesh, F.; Kizhakkethottam, J.J. Texture-based Estimation of Age and Gender fromWild Conditions. Procedia Technol. 2016, *24*, pp. 1349–1357.

[5] Kohli, S., Prakash, S., Gupta, P., Age Estimation Using Active Appearance Models and Ensemble of Classifiers with Dissimilarity-Based Classification. In Advanced Intelligent Computing; Springer: Berlin/Heidelberg, Germany, 2011; pp. 327–334.

[6] Angulu, R.; Tapamo, J.-R.; Adewumi, A.O., Age estimation via face images: A survey. EURASIP J. Image Video Process., *42*, 2018.

[7] Yamashita, R.; Nishio, M.; Do, R.K.G.; Togashi, K. Convolutional neural networks: An overview and application in radiology. *Insights into Imaging*, 2018, *9*, pp. 611–629.

[8] He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016; pp. 770–778, doi: 10.1109/CVPR.2016.90.

[9] Sigal, L., Human Pose Estimation. In: Ikeuchi, K. (eds) Computer Vision. Springer,2021, Cham. https://doi.org/10.1007/978-3-030-63416-2_584.

[10] Maggiori E. , Tarabalka Y. , Charpiat G. and Alliez P., "Can Semantic Labeling Methods Generalize to Any City? The Inria Aerial Image Labeling Benchmark". IEEE International Geoscience and Remote Sensing Symposium (IGARSS). 2017, pp. 3226-3229, doi: 10.1109/IGARSS.2017.8127684..

[11] Insafutdinov E. , Pishchulin L. , Andres B., Andriluka M. , and Schiele B. . Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In ECCV, 2016, pp. 34–50.

[12] Zhou X., Wang D., and Krahenbuhl P. Objects as points., In arXiv preprint arXiv:1904.07850, 2019.

[13] Chen Y., Wang Z., Peng Y., Zhang Z., Yu G., and Sun, J., Cascaded pyramid network for multi-person pose estimation. In IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7103-7112, doi: 10.1109/CVPR.2018.00742.

[14] Kocabas M., Karagoz S., and Akbas E., Multiposenet: Fast multi-person pose estimation using pose residual network. In ECCV. 2018, pp.1-17.

[15] Papandreou G. , Zhu T. , Chieh Chen L. , Gidaris S., Tompson J., Murphy K. , Person Pose Estimation and Instance Segmentation with a Bottom-Up, Part-Based, Geometric Embedding Model. Computer Vision and Pattern Recognition, CoRR, abs/1803.08225v1, http://arxiv.org/abs/1803.08225, 2018.

[16] Chung J. , Ong L. and Leow M., Comparative Analysis of Skeleton-Based Human Pose Estimation, Future Internet, 2022, pp. 1-1914(12), *380*; https://doi.org/10.3390/fi14120380

[17] Zielinski E., Live Perception and Real Time Motion Prediction with Deep Neural Networks and Machine Learning. Master's thesis, Harvard University Division of Continuing Education, 2021, 73 pages.

[18] Masud M, Muhammad G, Alhumyani H, Alshamrani SS, Cheikhrouhou O, Ibrahim S, Hossain MS., Deep learning-based intelligent face recognition in iot-cloud environment. Comput Commun. 2020; PP. 215–222, *152*.

[19] Xiao, J.; Wang, J.; Cao, S.; Li, B. Application of a Novel and Improved VGG-19 Network in the Detection of Workers Wearing Masks. J. Phys. Conf. Ser., 2020, *1518*.

[20] Sandler M. , Howard A. , Zhu M. , Zhmoginov A. and Chieh Chen L., Inverted Residuals and Linear Bottlenecks: Mobile Networks for Classification, Detection and Segmentation, Computer Vision and Pattern Recognition, CoRR, abs/1801.04381, 2018, http://arxiv.org/abs/1801.04381

[21] Howard A. , Zhu M., Chen B., Kalenichenko D., Wang W., Weyand T., Andreetto M. and Adam H., MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications,CoRR, abs/1704.04861v1, http://arxiv.org/abs/1704.04861,2017

[22] Orozco-Arias S., Piña J., Tabares-Soto R., Castillo-Ossa L., Guyot R. ,and Isaza G., Measuring Performance Metrics of Machine Learning Algorithms for Detecting and Classifying Transposable Elements, Processes, 2020, *8* (6), https://doi.org/10.3390/pr8060638, pp. 1-19.