A Deep CNN Model for Inferring 3D Human Body Shapes Using Front and Side Images

Elena ÁLVAREZ DE LA CAMPA CRESPO, Bernhard SPANLANG
Virtual Bodyworks S.L., Barcelona Health Hub, Barcelona, Spain

https://doi.org/10.15221/21.32

Abstract

Immersive Virtual Reality (IVR) studies indicate that there is some level of the brain that does not distinguish between reality and virtual reality. In this context, a self avatar embodied from first person perspective brings a significant and lasting change to the user. IVR is therefore widely used in research and for psychological and physiological health rehabilitation. We use IVR in a wide range of areas in pain and mobility, emotional health, diversity equity and inclusion, to rehabilitate domestic violence offenders and to promote healthier lifestyles among obese people (Fig. 1). A remaining challenge is to accurately and efficiently create avatars with body shapes and appearance that closely match those of the real user's bodies. This is owing to the huge differences in human body forms, the reduction of the complex human shape by body scanning technology and the complexity of acquiring accurate body measurements.

The primary objective of this work was to construct a cost-effective and accurate model to infer the 3D shape from a front and side image of a person taken with a smartphone. To achieve this, we used a fully morphable human body model to change the body shape using a set of body shape modifying parameters. We create a dataset of thousands of computer generated front and side images varying the shape modifiers of the morphable model. We then train a convolutional neural network (CNN) using that dataset (Fig. 3).

Our approach efficiently infers 3D human body shapes from a person's front and side image generating an accurate representation of a person. We made preliminary tests using a set of 10 body scans with known measurements, creating computer generated front and side images of the scans and using these images as input to the CNN and to compare the resulting body shape with the original 3D body scan. Our results demonstrate the effectiveness of the designed approach (Fig. 5, Fig. 6).

Our proposed model enables us to create a fully movable avatar that can be embodied in IVR from a front and side smartphone photo in a fully automated way. The same inferred shape modifiers can also be used on the clothing of the avatar to enable us to dress the avatar. Although a larger comparative study needs to be performed before the use of this approach can be routinely recommended, we believe that the convenience and ease-of-use of this model will contribute to increase the reach of VR tools with look-alike avatars also in clinical settings.

Keywords: 3D Body Scanning, Body Measurement, 3D Body Shape, Virtual Reality, Embodiment, Deep Learning, Machine Learning, Health Application, Convolutional Neural Networks, Computer Vision

1. Introduction

In cognitive neuroscience, a very fundamental question is how the brain represents our own body [1]. Here, when many of us think of our own body representation, we see it as a stable and unchanging depiction. However, several studies have induced in different people the illusion of body ownership over external objects that were not part of the human body at all. This refutes the immutability of our own body representation and suggests that this representation can indeed be very malleable [1-3].

Additionally, it has been shown that the illusion of body ownership can also spread and be induced throughout the whole body [4-6]. In this line of thought, several studies have shown that in order to induce this body ownership across the whole body, a first-person perspective of the artificial body is key [4-5,7]. This whole-body ownership illusion from a first-person perspective is described as the sensation of possessing an artificial body, in which the real body is substituted as the origin of the perspective sensations.

Nowadays, this so-called body ownership illusion can be induced with immersive virtual reality (IVR) systems that allow users to substitute a person's real body for a virtual representation. IVR users have the sensation of being "embodied" in that virtual body [5,8]. Furthermore, with the use of IVR we can act on the virtual body seen from a first-person perspective and experimentally manipulate the multisensorial integration in a highly controlled way. For example, you can manipulate the representation of the body in terms of structure, shape, size, and color in a way that contrasts sharply with our image of our own body [9-12]. In addition, it has been shown that by manipulating the properties of the virtual body, it is possible to influence the physiological response of the real body [13-14], and may also modulate behavioral responses of the subjects [15-16], creating a substantial and lasting change to the user.

For this reason, immersive virtual reality has been found to have many potential uses in psychotherapy, rehabilitation, behavioural neuroscience [17-19], and even consciousness research [20]. IVR is therefore widely used in physical and mental health research and rehabilitation. This is the case, for instance, of our projects and developments. We use break throughs in neuroscience and virtual reality research to change society in a wide range of areas: pain and mobility, emotional health, diversity equity and inclusion, to rehabilitate domestic violence offenders and to promote healthier lifestyles among obese people (Fig. 1).

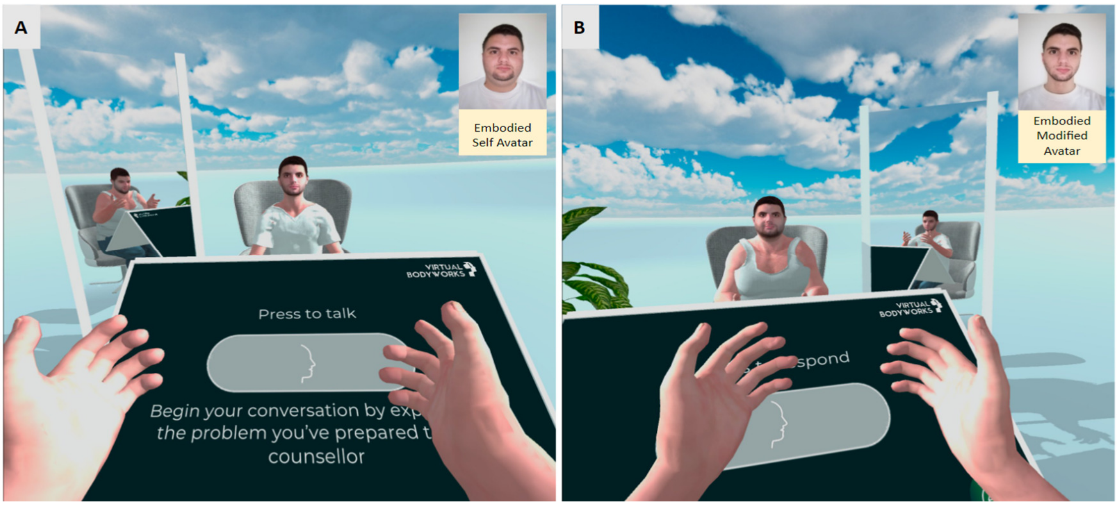

Fig. 1. A virtual scenario of our conVRself specially adapted to promote healthier lifestyles among obese people. (A) An overview of the scene from just behind the viewpoint of the participant who can see him/herself in the mirror and the healthier virtual representation of him/herself across the table. (B) From the virtual body of the healthier virtual representation of him/herself the participant can see his/her reflection in the mirror and the representation of him/herself across the table.

One of the key problems in merging VR and neurosciences applications originates from the importance of having a good self-body representation in an IVR experience. Some models already exist today that deal with 3D body scan using booth scanners, portable scanners, or mobile applications converting 2D photos to 3D models. However, most of them show that an area that is still pending is to accurately and efficiently create avatars with body shapes and appearance that closely match those of the real user's body. This is owing to the huge variety of human body forms, the reduction of the complex human shape by body scanning technology and the complexity of acquiring accurate body measurements.

In this paper, we constructed a cost-effective and accurate model to infer the 3D human body shape from a front and side image of a person taken with a smartphone using Convolutional Neural Networks (CNN). Our proposed model, which efficiently infers human silhouettes, enables us to create a fully movable and clothed avatar that can be embodied in IVR.

2. Materials and Methods

In this section, we introduced our dataset using a fully morphable human body model, followed by its preprocessing methodology and training, validation, and testing criteria. Then, we describe the model architecture of our proposed approach and the validation process.

2.1. Data Collection

We collected overall 4262 computer generated front and side images using a fully morphable human body model. This fully morphable human body model is a fully animatable avatar including body skeleton and facial morph targets adapted to the *SMPL* shape parameters. The Skinned Multi-Person Linear model (*SMPL*) [21] is a skinned vertex-based model that accurately represents a wide variety of body shapes in natural human poses while modifying only 10 shape parameters (blendshapes). This simple formulation enables us training our model using a relative large number of front and side images of different people with different body sizes.

For training purposes, we need a dataset which has images and labels attached to it. Each front and side image has as label its own combination of blendshape values.
The original image resolution was 1080 x 1080 pixels. The main objective related to this dataset is the automatic inference of the combination of blendshape values, most importantly that the body shape closely matches that of the real user's body.

2.2. Preprocessing

The dataset used in this paper contains numerous front and side images of different body sizes maintaining the same whole body pose. However, these conditions are difficult to maintain, as person position and rotation may be different in future front and side images. Moreover, sometimes some body parts (arms, feet, head, hair) are redundant for the training process and can cause mismatches when predicting the bodyshape of a person. Other redundant areas are the entire background of the image outside the person's silhouette. As these wrong postures, differences on the person position, unnecessary body parts or leftover pixels may adversely affect the training process of the CNN model, we decided to apply a body part segmentation tool to process each front and side 1080 x 1080 image.
There are many deep learning architectures which could be used to solve the instance segmentation problem [22], however, in this work we used an improved baseline from *DensePose* [23], a project from *Detectron2* [24] which is a Facebook AI Research's next generation library that provides state-of-the-art detection and segmentation algorithms. Its goal is to assign semantic labels (e.g., person, sheep, airplane and so on) and, once a person is identified, then mapping all human pixel of an RGB image to the 3D surface of the human body and to annotate the different human body parts (trunk, arm, legs, etc).
After having tested different models, we have focused on using the one that gave us the best performance when it comes to correctly separate body parts of each person: *densepose_rcnn_R_101 _FPN_DL_s1x.yaml*, an improved baseline of *DensePose* with a *DeepLabV3* Head. The applied model uses an improved training schedule, Panoptic FPN head [25] and DeepLabV3 head [26]..
Following this procedure, front and side person images are first human detected and segmented into different body parts. Next, coloured images are converted into greyscale images. Then, images are downscaled to the desired dimension of 128 x 128 pixels. Finally, front and side image dataset are independently processed. Front dataset contains trunk and leg as separated images, while side dataset contain trunk and leg as a whole image.
The example of original and body segmented and processed images are shown in Fig. 2.

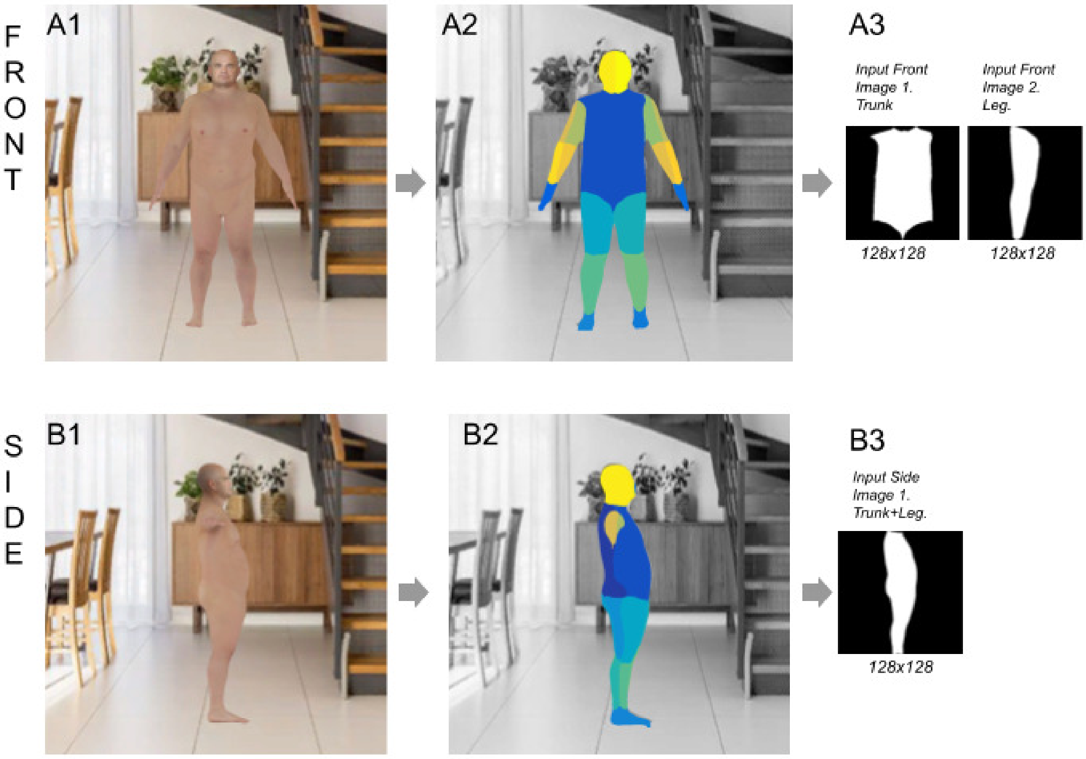
F R O N T
A1
A2
A3
Input Front Image 1. Trunk
Input Front Image 2. Leg.
128x128
128x128
S I D E
B1
B2
B3
Input Side Image 1. Trunk+Leg.
128x128

Fig. 2. The example of original front (A1) and side (B1) computer generated images, human detection and body part segmentation (A2,B2), and independent processing of front and side images (A3,B3).

2.3. Training Criteria

For the model construction, we selected 80% of images for training and the remaining 20% for testing purposes, with similar percentage of different body size images. In this way, 3409 images were used for training whereas the remaining 853 images were kept for testing the model. Following [27], we used 5-fold cross-validation on training images, which means that 2727 images were used for training and 682 images for validation purpose. The standard protocol [28-29] is based on partitioning the original data in N subsets and execute the algorithm N times. During each run, N-1 subsets are used as training data and the remaining subset as testing data in which we will estimate the model's performance. This approach gives a much larger training set for each test, which means that the algorithm will have enough data to learn from without overfitting. Once each subset (of variants) has been set aside as a test set, we can obtain the model performance, by averaging that of the test sets.

2.4. CNN architecture

Convolutional Neural Networks (CNN) [30] have applications in image and video recognition, recommender systems, image classification, image segmentation, medical image analysis, natural language processing, brain-computer interfaces, and financial time series. CNNs are regularized versions of multilayer perceptrons. Multilayer perceptrons usually mean fully connected networks, that is, each neuron in one layer is connected to all neurons in the next layer. Regularization is the process of adding information in order, for instance, to prevent overfitting. They are based on the shared-weight architecture of the convolution kernels or filters that slide along input features and provide translation equivariant responses known as feature maps.

As one of the state-of-the-art algorithms in the field of computer vision, CNN can effectively extract spatial-related high level features. Different from the manual setting parameters of traditional spatial feature extraction methods, the CNN-based method can automatically learn the spatial-related parameters layer by layer [31-32].

In this model (Fig. 3), there are 3 CNN blocks, and each block consists of 2 convolution layers, 2 max-pooling layers and a fully-connected layer. The *Relu* activation function is used to remove negative values from the feature map because there can not be negative values for any pixel value. Stride(1,1) used and padding is also 1. Then, there is a final block that consists of a fully-connected layers that merges extracted features from the three input images.

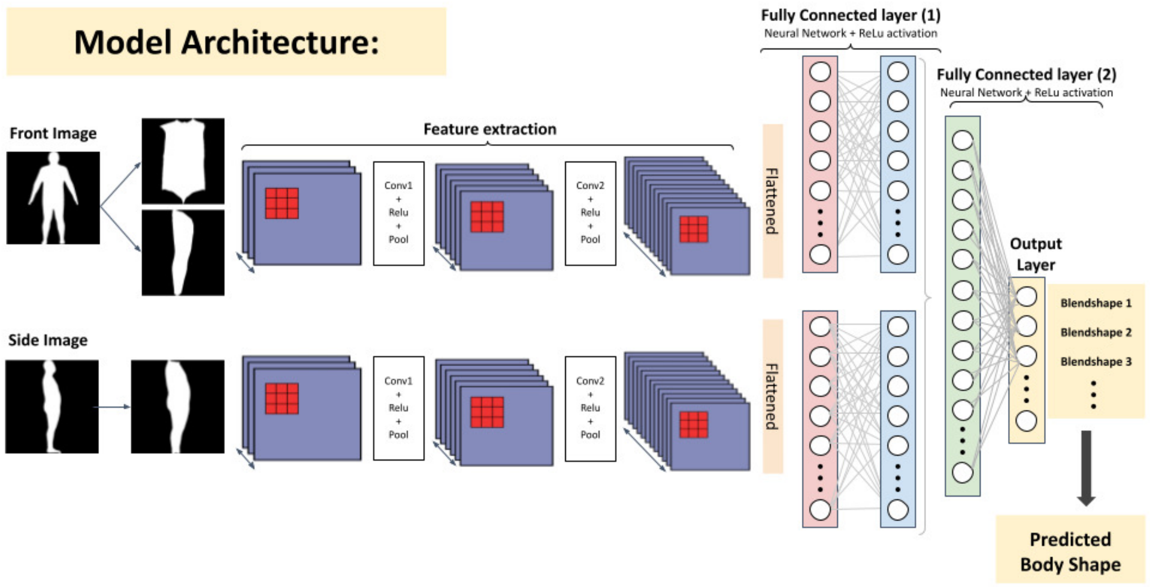

Fig. 3. Architecture from the CNN model to infer 3D body shapes using front and side images.

2.5. Experimental environment

We implemented all the experiments related to this article by using Python 3.7.9 along with TensorFlow 1.15.4, Torch 1.4.0 and TorchVision 0.5.0 installed on a standard PC with Nvidia GeForce GTX 1070 graphics processing unit (GPU) support. Moreover, this PC has a RAM capacity of 16.0 GB and holds a 3.60 GHz Intel(R) Core(TM) i7-7700K processor.

2.6. Evaluation Metrics

Until now, we have spoken about performance in an abstract way. At a practical level, however, there are different options to measure it, that is, to estimate the success rate of a method. Any performance measure is based on four fundamental quantities obtained after applying our predictor to any mixture of known variants [33]: TP (true positives); TN (true negative); FN (false negative); FP (false positive). These four quantities are then combined to produce single measures of performance. As a combination of these four quantities, accuracy (ACC) is frequently employed as is considered to be more informative than other measures [34]. ACC evaluates the correctness of a model, and is the ratio of the number of images accurately classified out of the total number of testing images.

$$Accuracy\ (ACC) = \frac{TP + TN}{TP + FP + TN + FN}$$

As it is advised to describe the performance of a predictor providing several measures, we also defined an evaluation metric suitable to be used for a regression. Unlike classification, accuracy in a regression model is slightly harder to illustrate. It is impossible for you to predict the exact value, but rather how close your prediction is against the real value. We used the Mean Square Error (MSE) as an absolute measure of the goodness for the fit [35].

$$Mean\ Square\ Error\ (MSE) = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \hat{y}_i\right)^2$$

MSE is calculated by the sum of square of prediction error which is real output minus predicted output and then divide by the number of data points. It gives you an absolute number on how much your predicted results deviate from the actual number. Together with the model accuracy, it gives us a real number to compare against other model results and help us select the best regression model.

2.7. Hyperparameter Tuning

Neural networks have the ability to learn complex connections between their inputs and outputs. Some of these connections, however, may be the consequence of sampling noise, thus they may appear during the training phase but not in the real test dataset. In this case, overfitting can occur and therefore can degrade the prediction performance of a deep learning algorithm [36]. Therefore, we followed the procedure of tuning the hyperparameters to obtain the overall performance of the proposed model [37]. The methods used to select the best hyperparameters are: First, select the mean squared error as the loss function for the multi-class regression problem. Then, during training, Adam's algorithm (adaptive momentum estimation) was used to perform a 500 epoch optimization. During this period, we tried several learning rates (0.10-0.00001) and different batch sizes (1, 16, 20, 32, 50, 6, 100, 150, 300, 500, 1000). During the model training, our main goal was to minimize the overall difference between training loss and validation loss. Batch size 20 was found to work well with a learning rate of 0.001. Additionally, we used a dropout rate of 0.2 to avoid the model from overfitting during training. We then used the 5-fold cross-validation method based on its minimal validation loss by using a 5-fold cross validation approach. Finally, we used these weights for the multi-class regression on the test dataset.

2.8. “Real-World” External Test Data

In this work, we used CAESAR data [38], which contains three-dimensional (3D) whole-body scans, to additionally test our constructed model. We demonstrate the ability and the effectiveness of the reported deep learning approach in the “real-world” data by comparing model predicted bodyshapes with scanned CAESAR body shapes. We used and compared 4 discrete measurements as a description of the body shape: waist, hip, chest and leg.

3. Results and Discussion

Convolutional Neural Networks (CNN) are widely used for image recognition and classification. In this work, the performance of CNN were evaluated for inferring human body shapes from a person’s front and side image. Table 1 shows the evaluation metrics of the train set, validation set and test set of the CNN model. To consider a model as the best model, it must have performed well in the training data set and the validation data set [39]. In this sense, results shown in Table 1 were found to be from the best model.

Table 1. Evaluation metrics of the best CNN model.

Data Set	Evaluation Metrics	CNN Model
Train	MSE Loss	0,000191
	Accuracy	0,999617
Validation	MSE Loss	0,000265
	Accuracy	0,993487
Test	MSE Loss	0,000314
	Accuracy	0,987768

As we can see in the previous table, in the computer generated front and side images using a fully morphable human body model, the CNN model was found to reach 98,78% accuracy on the test set. This high performance of the model on the test set validates the generalization ability of the model.
During the model construction, the variation trend of training loss almost overlaps validation loss (Fig. 4A). At the same time, the variation trend of training accuracy is additionally according to that of validation accuracy (Fig. 4B). This situation shows that the model does not have an overfitting problem with the parameters chosen during the training process. These results show that the proposed architecture can distinguish the matched images within the input well and thus efficiently infer 3D human body shapes from a person's front and side image, generating an accurate representation of a person.

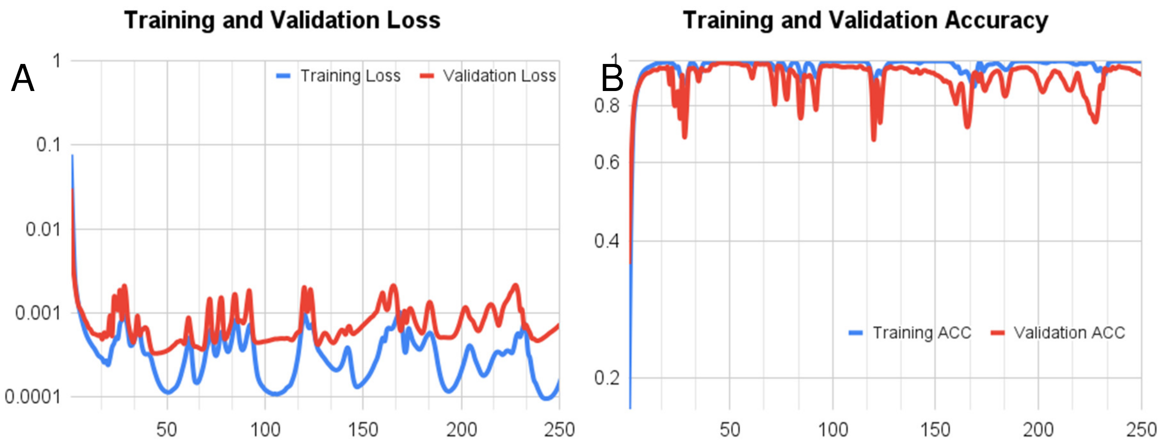

Fig. 4. Training and Validation Loss (A) and Accuracy (B) of the CNN model during the model training for inferring human body shape from front and side images.

Additionally, we also used an external validation dataset: the CAESAR data, which contains three-dimensional (3D) whole-body scans. After generating front and side images of each CAESAR sample, we tested the CNN model on them and compared 4 discrete body shape measurements (waist, hip, chest and leg circumferences). Each body measurement was calculated for each 3D human model separately. The mean of the differences between CNN model prediction and CAESAR scan per each body measurement is shown in Fig. 5A, table on the left side. According to the table, and the small difference in centimetres, it can be said that predicted measurements by the CNN model are on average highly similar to the real CAESAR scans. In Fig. 5B, the similarity between the predicted body shapes and CAESAR scans are also depicted.

A

Measurement (cm)	Mean differences between CEASAR and Predicted
Waist	0,8
Hip	2,1
Chest	1,6
Leg	1,2

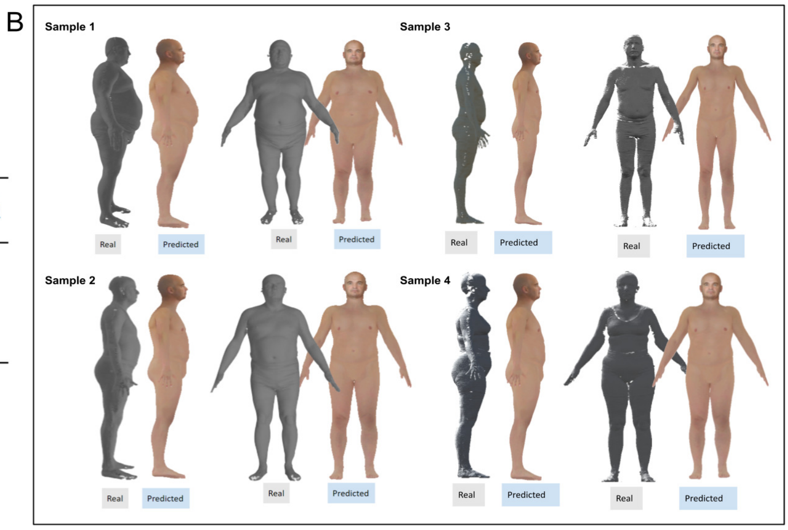

Figure 5. The table on the left shows the mean measurement differences from CAESAR scans and Predicted Body Shape. Images on the right show 4 from 10 tested samples. Grey avatars are real scanned humans, while textured avatars are the predicted body shapes.

In order to complete the above analysis on model viability, we tested our CNN model with a real person using a front and side image (Fig. 6).

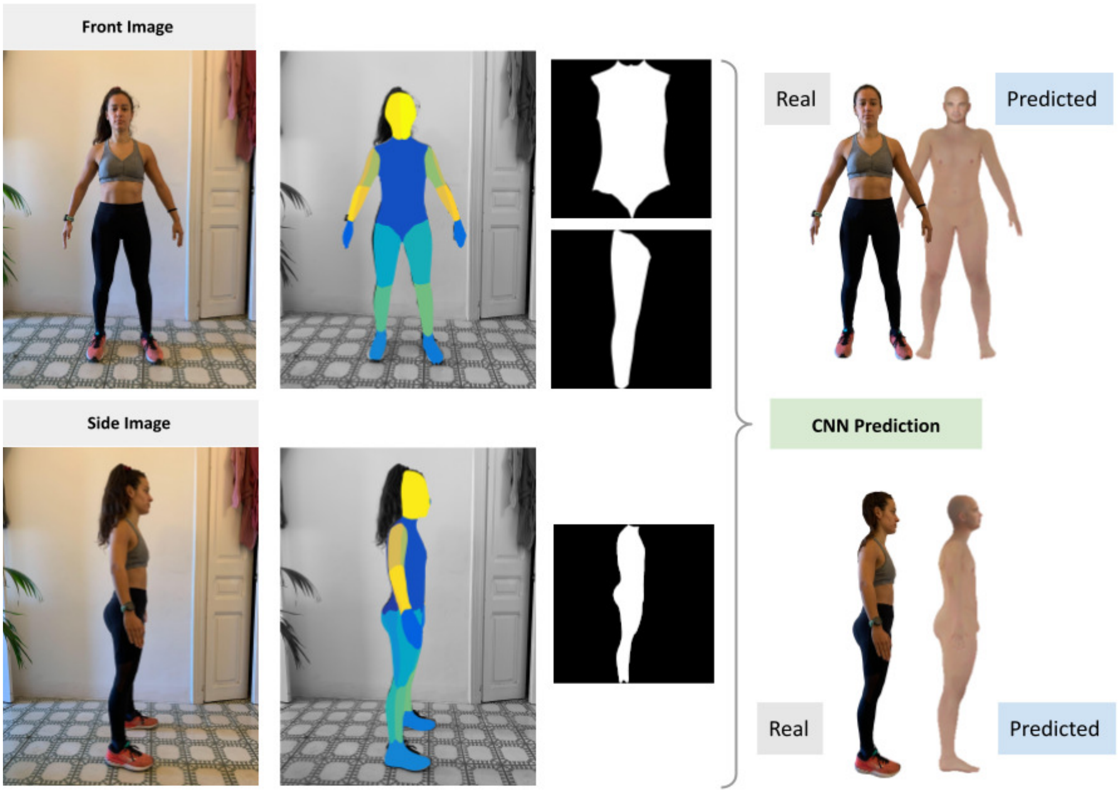

Fig. 6. Testing the Constructed Model to infer 3D Human Body Shape with a real sample. Front and side images were taken from a participant. Then, extracting the silhouette and segmenting the different body parts. Finally, generate black and white 128x128 images and test the CNN model. Real and predicted 3D body shapes are shown on the right side of the figure.

In summary (Fig. 5 and Fig. 6), our results demonstrate the effectiveness of the presented approach. Our CNN model accurately infers 3D human body shapes from a person's front and side image generating an accurate representation of a person and creates a fully movable avatar that can be embodied and used in IVR. Moreover, the same inferred shape modifiers can also be used on the clothing of the avatar to enable us to dress the avatar.

4. Conclusions

In this paper, we have proposed a novel model that efficiently infers 3D human body shapes from a person's front and side image, generating an accurate representation of a person. The proposed method has the following characteristics: 1) overcomes the singularity and the variety on the different human body shapes in both men and women; 2) the CNN-based spatial feature extraction technique overcomes the inefficient and unaccurate process of manually selecting the avatar's body shape; 3) combining CNN and Semantic Image Segmentation leads to excellent classification accuracies; 4) it creates a fully morphable avatar with a series of shape modifiers that allows to change the body shape of an embodied avatar at run time inside any IVR experience; 5) The same inferred shape modifiers can also be used on the clothing of the avatar to enable us to dress the avatar; 6) the process from a person's image to a fully movable avatar is done in a fully automated way; 7) we believe that the convenience and ease-of-use of this model will contribute to increase the reach of VR tools with look alike avatars also in clinical settings.

However, the proposed method can still be revised in some aspects. For instance, a larger comparative study needs to be performed before the use of this approach can be routinely recommended. Also, we would like the final approach to be extensible to other human models, not only applicable to *SMPL* models. Therefore, our future work will focus first on improving the performance validation and, second, on how to select the optimal approach and existing human models in order to obtain a fully morphable avatar with the best 3D body shape inference.

5. Acknowledgments

This research was funded under the project "Self Conversation in Virtual Reality Embodiment to Enhance Healthier Lifestyles Among Obese People" (SOCRATES) H2020-EU.2.1.1. - Leadership in enabling and industrial technologies - Information and Communication Technologies (ICT) Grant agreement 951930.

6. References

[1] M. Slater and M. V. Sanchez-Vives, "Enhancing Our Lives with Immersive Virtual Reality," Frontiers in Robotics and AI, vol. 3, Dec. 2016, doi: 10.3389/frobt.2016.00074.

[2] M. Botvinick and J. Cohen, "Rubber hands 'feel' touch that eyes see," Nature, vol. 391, no. 6669, pp. 756–756, Feb. 1998, doi: 10.1038/35784.

[3] K. C. Armel and V. S. Ramachandran, "Projecting sensations to external objects: evidence from skin conductance response," Proceedings of the Royal Society of London. Series B: Biological Sciences, vol. 270, no. 1523, pp. 1499–1506, Jul. 2003, doi: 10.1098/rspb.2003.2364.

[4] V. I. Petkova, M. Khoshnevis, and H. H. Ehrsson, "Mannequin Body Illusion Measure," PsycTESTS Dataset, 2011, doi: 10.1037/t31408-000.

[5] M. Slater, B. Spanlang, M. V. Sanchez-Vives, and O. Blanke, "First Person Experience of Body Transfer in Virtual Reality," PLoS ONE, vol. 5, no. 5, p. e10564, May 2010, doi: 10.1371/journal.pone.0010564.

[6] A. Maselli and M. Slater, "The building blocks of the full body ownership illusion," Frontiers in Human Neuroscience, vol. 7, 2013, doi: 10.3389/fnhum.2013.00083.

[7] V. I. Petkova and H. H. Ehrsson, "If I Were You: Perceptual Illusion of Body Swapping," PLoS ONE, vol. 3, no. 12, p. e3832, Dec. 2008, doi: 10.1371/journal.pone.0003832.

[8] M. Slater, "Place illusion and plausibility can lead to realistic behaviour in immersive virtual environments," Philosophical Transactions of the Royal Society B: Biological Sciences, vol. 364, no. 1535, pp. 3549–3557, Dec. 2009, doi: 10.1098/rstb.2009.0138.

[9] K. Kilteni, R. Groten, and M. Slater, "The Sense of Embodiment in Virtual Reality," Presence: Teleoperators and Virtual Environments, vol. 21, no. 4, pp. 373–387, Nov. 2012, doi: 10.1162/pres_a_00124.

[10] K. Kilteni, J.-M. Normand, M. V. Sanchez-Vives, and M. Slater, "Extending Body Space in Immersive Virtual Reality: A Very Long Arm Illusion," PLoS ONE, vol. 7, no. 7, p. e40867, Jul. 2012, doi: 10.1371/journal.pone.0040867.

[11] D. Banakou, R. Groten, and M. Slater, “Illusory ownership of a virtual child body causes overestimation of object sizes and implicit attitude changes,” Proceedings of the National Academy of Sciences, vol. 110, no. 31, pp. 12846–12851, Jul. 2013, doi: 10.1073/pnas.1306779110.

[12] T. C. Peck, S. Seinfeld, S. M. Aglioti, and M. Slater, “Putting yourself in the skin of a black avatar reduces implicit racial bias,” Consciousness and Cognition, vol. 22, no. 3, pp. 779–787, Sep. 2013, doi: 10.1016/j.concog.2013.04.016.

[13] M. Martini, D. Perez-Marcos, and M. V. Sanchez-Vives, “What Color is My Arm? Changes in Skin Color of an Embodied Virtual Arm Modulates Pain Threshold,” Frontiers in Human Neuroscience, vol. 7, 2013, doi: 10.3389/fnhum.2013.00438.

[14] D. Banakou, P. D. Hanumanthu, and M. Slater, “Virtual Embodiment of White People in a Black Virtual Body Leads to a Sustained Reduction in Their Implicit Racial Bias,” Frontiers in Human Neuroscience, vol. 10, Nov. 2016, doi: 10.3389/fnhum.2016.00601.

[15] S. A. Osimo, R. Pizarro, B. Spanlang, and M. Slater, “Conversations between self and self as Sigmund Freud—A virtual body ownership paradigm for self counselling,” Scientific Reports, vol. 5, no. 1, Sep. 2015, doi: 10.1038/srep13899.

[16] S. Seinfeld et al., “Offenders become the victim in virtual reality: impact of changing perspective in domestic violence,” Scientific Reports, vol. 8, no. 1, Feb. 2018, doi: 10.1038/s41598-018-19987-7.

[17] M. J. Tarr and W. H. Warren, “Virtual reality in behavioral neuroscience and beyond,” Nature Neuroscience, vol. 5, no. S11, pp. 1089–1092, Oct. 2002, doi: 10.1038/nn948.

[18] M. Martini, “Real, rubber or virtual: The vision of ‘one’s own’ body as a means for pain modulation. A narrative review,” Consciousness and Cognition, vol. 43, pp. 143–151, Jul. 2016, doi: 10.1016/j.concog.2016.06.005.

[19] G. Riva, B. K. Wiederhold, and F. Mantovani, “Neuroscience of Virtual Reality: From Virtual Exposure to Embodied Medicine,” Cyberpsychology, Behavior, and Social Networking, vol. 22, no. 1, pp. 82–96, Jan. 2019, doi: 10.1089/cyber.2017.29099.gri.

[20] M. V. Sanchez-Vives and M. Slater, “From presence to consciousness through virtual reality,” Nature Reviews Neuroscience, vol. 6, no. 4, pp. 332–339, Apr. 2005, doi: 10.1038/nrn1651.

[21] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, “SMPL,” ACM Transactions on Graphics, vol. 34, no. 6, pp. 1–16, Nov. 2015, doi: 10.1145/2816795.2818013.

[22] A. M. Hafiz and G. M. Bhat, “A survey on instance segmentation: state of the art,” International Journal of Multimedia Information Retrieval, vol. 9, no. 3, pp. 171–189, Jul. 2020, doi: 10.1007/s13735-020-00195-x.

[23] R. A. Guler, N. Neverova, and I. Kokkinos, “DensePose: Dense Human Pose Estimation in the Wild,” Jun. 2018, Accessed: Aug. 31, 2021. [Online]. Available: http://dx.doi.org/10.1109/cvpr.2018.00762.

[24] Yuxin Wu and Alexander Kirillov and Francisco Massa and Wan-Yen Lo and Ross Girshick, “Detectron2,” [Online]. Available: https://github.com/facebookresearch/detectron2.

[25] A. Kirillov, K. He, R. Girshick, C. Rother, and P. Dollar, “Panoptic Segmentation,” Jun. 2019, Accessed: Aug. 31, 2021. [Online]. Available: http://dx.doi.org/10.1109/cvpr.2019.00963.

[26] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, “DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 40, no. 4, pp. 834–848, Apr. 2018, doi: 10.1109/tpami.2017.2699184.

[27] Andrew W. Moore, “Cross-validation for detecting and preventing overfitting,” 2011, [Online]. Available: https://www.cs.cmu.edu/~./awm/tutorials/overfit10.pdf.

[28] V. G. Krishnan and D. R. Westhead, “A comparative study of machine-learning methods to predict the effects of single nucleotide polymorphisms on protein function,” Bioinformatics, vol. 19, no. 17, pp. 2199–2209, Nov. 2003, doi: 10.1093/bioinformatics/btg297.

[29] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, 2016.

[30] Y. Chen, H. Jiang, C. Li, X. Jia, and P. Ghamisi, “Deep Feature Extraction and Classification of Hyperspectral Images Based on Convolutional Neural Networks,” IEEE Transactions on Geoscience and Remote Sensing, vol. 54, no. 10, pp. 6232–6251, Oct. 2016, doi: 10.1109/tgrs.2016.2584107.

[31] A. Farooq, J. Hu, and X. Jia, “Weed Classification in Hyperspectral Remote Sensing Images Via Deep Convolutional Neural Network,” Jul. 2018, Accessed: Aug. 31, 2021. [Online]. Available: http://dx.doi.org/10.1109/igarss.2018.8518541.

[32] L. Gomez-Chova, D. Tuia, G. Moser, and G. Camps-Valls, “Multimodal Classification of Remote Sensing Images: A Review and Future Directions,” Proceedings of the IEEE, vol. 103, no. 9, pp. 1560–1584, Sep. 2015, doi: 10.1109/jproc.2015.2449668.

[33] S. V. Stehman, “Selecting and interpreting measures of thematic classification accuracy,” Remote Sensing of Environment, vol. 62, no. 1, pp. 77–89, Oct. 1997, doi: 10.1016/s0034-4257(97)00083-7.

[34] P. Baldi, S. Brunak, Y. Chauvin, C. A. F. Andersen, and H. Nielsen, “Assessing the accuracy of prediction algorithms for classification: an overview,” Bioinformatics, vol. 16, no. 5, pp. 412–424, May 2000, doi: 10.1093/bioinformatics/16.5.412.

[35] “Mean Squared Error (MSE).” https://www.probabilitycourse.com/chapter9/9_1_5_mean_squared_error_MSE.php (accessed Aug. 31, 2021).

[36] X. Ying, “An Overview of Overfitting and its Solutions,” Journal of Physics: Conference Series, vol. 1168, p. 022022, Feb. 2019, doi: 10.1088/1742-6596/1168/2/022022.

[37] U. Michelucci, “Hyperparameter Tuning,” in Applied Deep Learning, Berkeley, CA: Apress, 2018, pp. 271–322.

[38] K. M. Robinette and H. Daanen, “Lessons Learned from Caesar: A 3-D Anthropometric Survey,” Defense Technical Information Center, Fort Belvoir, VA, Jan. 2003. Accessed: Aug. 31, 2021. [Online]. Available: http://dx.doi.org/10.21236/ada430674.

[39] A. Taner, Y. B. Öztekin, and H. Duran, “Performance Analysis of Deep Learning CNN Models for Variety Classification in Hazelnut,” Sustainability, vol. 13, no. 12, p. 6527, Jun. 2021, doi: 10.3390/su13126527.