Deep Body Fitness

Fei Zhao, Chengcui Zhang, and Sheikh Abujar
Department of Computer Science
The University of Alabama at Birmingham
Birmingham, USA
{larry5, czhang02, sabujar}@uab.edu

***Abstract*—Traditional computer vision-based methods for estimating body fat percentage (BFP) rely on RGB images of the whole body, posing a risk of inadvertently leaking personal and private user information. Moreover, these methods often depend on hand-crafted features that are unreliable, highly customized, and computationally expensive. To address these challenges, this paper introduces a novel two-stage framework. In the first stage, we adopt a VGG19-UNet model to segment and extract only the body contours from RGB images, effectively removing irrelevant environmental content and ensuring that no re-identifiable details are disclosed. In the second stage, we implement spatial attention mechanisms to focus the model on crucial areas of body shapes. Additionally, the visual and demographic features are fused in an automated manner, enhancing the robustness of the model. Our experiments achieved a Root Mean Square Error (RMSE) of 4.13, representing a 29.5% enhancement from the state-of-the-art (SOTA) benchmark of 5.86.**

***Index Terms*—Body Fat Percentage, Multimodal Learning**

I. Introduction

Effective management and prevention of obesity are essential for improving health outcomes and reducing the healthcare burden worldwide. Traditional computer vision-based methods for estimating body fat percentage, such as those in [1] and [2], rely on handcrafted features that may be unreliable under varying lighting conditions or camera settings. Although modern deep learning models, e.g., [3], offer robust feature extraction from images, they typically require detailed whole-body RGB images, raising privacy concerns. To overcome these issues, we introduce a novel two-stage, deep learning-based multimodal approach that enhances the precision of BFP estimation. Our contributions are summarized as follows:

1) Enhanced Privacy: Our framework, relying solely on body contours extracted from RGB images, effectively protects user privacy. By avoiding the use of original RGB images that display identifiable personal details, e.g., face or body, we ensure privacy during the analysis process.

2) Multimodal Fusion Network: We propose an Attention-based network designed to efficiently extract deep features from body contour images. This approach eliminates the need for complicated, highly customized, and time-consuming computer vision methods to extract useful visual features for body fat prediction. Additionally, our model adeptly fuses visual features with demographic features in an automated manner, enhancing the accuracy and robustness of assessments.

This study was supported by resources from the P30 BIGDATA core grant NIH P30AR072583.

II. Methodology

Our work is a two-stage framework. The first stage involves precise segmentation of the body contour from the background and other non-essential elements, acting as a focused pre-processing step. The second stage utilizes this segmented body contour data in a multimodal network, integrating it with demographic information to improve the BFP estimation.

A. Body Contour Segmentation

We design a VGG19-based UNet model to perform precise segmentation of the body from input RGB images. We enhance the VGG19 [4] architecture with UNet's [5] efficient segmentation capabilities, enabling precise delineation of body contours. After segmentation, our method specifically extracts only the contours of the body from these segments. This critical step ensures that no re-identifiable information from the original image is used in the BFP estimation process. By focusing solely on the contours, we preserve user privacy while maintaining the integrity of the physiological features necessary for accurate BFP prediction.

B. Multimodal Fusion Nerual Network

We propose a ResNet152-based multimodal model to integrate segmented body contour data with demographic information. This process is divided into distinct phases:

1) Body Feature Extraction: Each segmented body contour image (back or side view), shown in Fig. 1, is initially processed through the ResNet152 subnetwork [6] for feature extraction. Subsequently, these extracted features are refined by a Spatial Attention mechanism [7], which consists of an AveragePooling layer, followed by a 3x3 Convolution (Conv) layer. This Conv layer generates an attention map that is then scaled by a Sigmoid function, emphasizing areas of interest while suppressing less relevant features. The attention-enhanced features proceed through an additional 3x3 Conv layer, which further sharpens the feature representation. A 2x2 MaxPooling layer then reduces the spatial dimensions of these feature maps, streamlining them for efficient processing. Finally, the compacted feature maps are flattened into 1D embeddings, preparing them for integration with demographic data in the fusion process.

2) Demographic Feature Extraction: The demographic data, including Age, Gender, Weight, BMI, and Race are concurrently processed through a 128-unit Dense layer with ReLU activation. A subsequent Dropout layer is utilized to

2835-5776/24/$31.00 ©2024 IEEE
DOI 10.1109/IRI62200.2024.00025

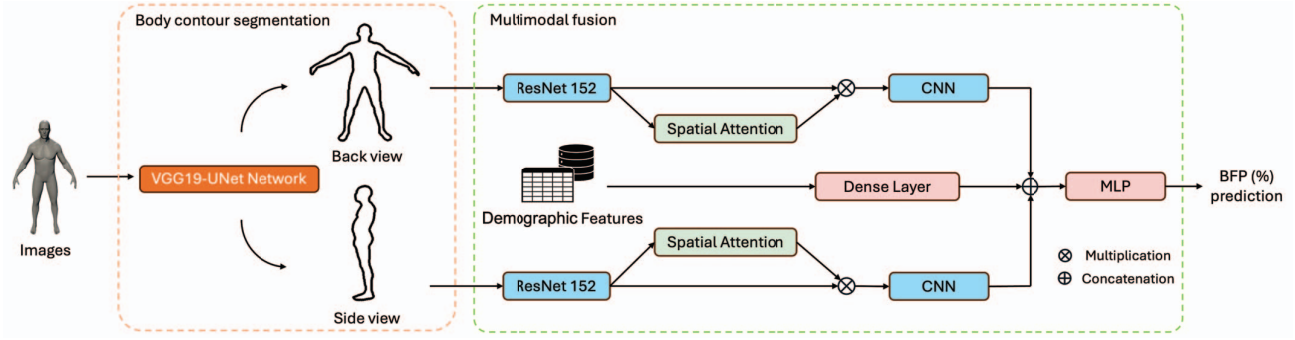

Fig. 1. Overall architecture for body fat percentage prediction

prevent overfitting. This process yields a refined embedding of demographic data, optimized for subsequent fusion.

3) Feature Fusion and BFP Estimation: Body features are concatenated with demographic embeddings and further processed through a Multi-layer Perceptron (MLP) with two Dense layers. The first layer, containing 256 units and ReLU activation, facilitates the nonlinear transformation and integration of the multimodal features. A subsequent Dropout layer with a rate of 0.4 helps mitigate overfitting. The final layer, featuring a single neuron, is designed to generate the BFP prediction. This structured fusion not only leverages critical visual features enhanced by the attention mechanism but also effectively combines demographic information to better predict body fat percentage.

III. Experiments and Results

We used the same demographic features in [1] and [2], as illustrated in Table I, and replace all the handcrafted computer vision features with the deep visual features aforementioned in our deep learning model. All body contour images were enhanced by an image dilation operation. Our dataset includes 322 samples, with 208 for training, 52 for validation, and 62 for testing. The model was trained with a learning rate of 0.001 and a batch size of 32, employing an early stopping mechanism capped at 100 epochs to prevent overfitting.

TABLE I
Visualization of some demographic data samples.

ID	Age	Sex	Race	Weight	BMI	BF%
2123	6	2	2	21	15.18	22.5
2007	14	1	2	47.2	20.27	17.5
4200	23	1	2	75	24.77	22.5

We compared our model against the SOTA NuSVR method [2] and other machine learning algorithms, including XGBoost, RandomForest, and AdaBoost, using the RMSE metric. The comparative RMSE results are detailed in Table II. Our model surpasses all compared models, illustrating its superior precision in estimating body fat percentage. The incorporation of the Spatial Attention mechanism significantly enhances the model's capabilities.

TABLE II
Final results on the test dataset

Model	RMSE Metric
NuSVR	5.86
XGBoost	5.36
RandomForest	4.88
AdaBoost	4.91
Our Model (Base Version)	4.44
Our Model (with Attention)	**4.13**

IV. Conclusion

Our research substantiates the efficacy of a multimodal deep-learning strategy in estimating body fat percentage (BFP). Future enhancements will focus on augmenting model performance by adopting more sophisticated fusion strategies and employing novel, lightweight network architectures. Such advancements are intended to improve the precision of BFP estimations and facilitate deployment on mobile platforms, thereby broadening the practical application of our model in real-world settings.

References

[1] O. Affuso, L. Pradhan, C. Zhang, S. Gao, H. W. Wiener, B. Gower, S. B. Heymsfield, and D. B. Allison, "A method for measuring human body composition using digital images," *PLoS One*, vol. 13, no. 11, p. e0206430, 2018.

[2] L. Pradhan, G. Song, C. Zhang, B. Gower, S. B. Heymsfield, D. B. Allison, and O. Affuso, "Feature extraction from 2d images for body composition analysis," in *2015 IEEE International Symposium on Multimedia (ISM)*, pp. 45–52, IEEE, 2015.

[3] S. S. Alves, E. F. Ohata, P. C. S. Junior, C. B. Barroso, N. M. Nascimento, L. L. Loureiro, V. Z. Bittencourt, V. L. M. C. Junior, A. R. da Rocha, and P. P. Rebouças Filho, "Sex-based approach to estimate human body fat percentage from 2d camera images with deep learning and machine learning," *Measurement*, vol. 219, p. 113213, 2023.

[4] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[5] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pp. 234–241, Springer, 2015.

[6] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[7] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 3–19, 2018.