Gender Recognition from 3D Shape Parameters

Giulia Martinelli[1 [0000−0003−3713−3053]], Nicola Garau[1 [0000−0001−7147−9109]], and Nicola Conci[1 [0000−0002−7858−0928]]

University of Trento, Via Sommarive, 9, 38123 Povo,Trento TN {giulia.martinelli-2, nicola.garau, nicola.conci}@unitn.it

Abstract. Gender recognition from images is generally approached by extracting the salient visual features of the observed subject, either focusing on the facial appearance or by analyzing the full body. In real-world scenarios, image-based gender recognition approaches tend to fail, providing unreliable results. Face-based methods are compromised by environmental conditions, occlusions (presence of glasses, masks, hair), and poor resolution. Using a full-body perspective leads to other downsides: clothing and hairstyle may not be discriminative enough for classification, and background cluttering could be problematic. We propose a novel approach for body-shape-based gender classification. Our contribution consists in introducing the so-called Skinned Multi-Person Linear model (SMPL) as 3D human mesh. The proposed solution is robust to poor image resolution and the number of features for the classification is limited, making the recognition task computationally affordable, especially in the classification stage, where less complex learning architectures can be easily trained. The obtained information is fed to an SVM classifier, trained and tested using three different datasets, namely (i) FVG, containing videos of walking subjects, (ii) AMASS, collected by converting MOCAP data of people performing different activities into realistic 3D human meshes, and (iii) SURREAL, characterized by synthetic human body models. Additionally, we demonstrate that our approach leads to reliable results even when the parametric 3D mesh is extracted from a single image. Considering the lack of benchmarks in this area, we trained and tested the FVG dataset with a pre-trained Resnet50, for comparing our model-based method with an image-based approach.

Keywords: gender recognition · body shape · parametric human body model.

1 Introduction

Gender recognition has a wide range of application areas, ranging from human-computer interaction to surveillance systems, as well as commercial developments with particular attention to retail analytics. For this task, the observation of the face is generally considered amongst the most relevant element of the body. However, there exists a large set of additional cues, which can be analyzed so as to infer the gender information. This includes, for example hairstyle, body shape, clothing, eyebrows, posture and gait, as well as vocal traits, based on the voice pitch. Such additional features allow for the

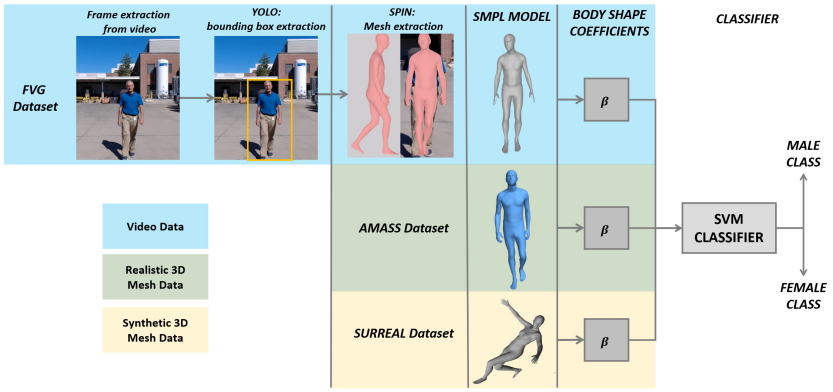

Fig. 1. Overview of the proposed pipeline. AMASS[12] and SURREAL[24] are characterized by SMPL[11] parametric mesh; the body shapes are therefore given. The FVG[26] dataset consists of videos of walking subjects. The parametric mesh is extracted using the SPIN[10] algorithm, from which the SMPL body shape coefficients are extracted and fed to the SVM classifier.

recognition through a multi-modal observation, exploring different dimensions, such as appearance, motion, and sound.

According to the information used for the classification, the existing gender recognition literature can be divided into two main categories: appearance and non-appearance-based approaches. The former leverages the features extracted from human physical appearance. These features can be static, denoting characteristics that are always present in an individual [6] (face, eyebrows, hand geometry), or dynamic [9], as body movement, activity recognition, or apparel information, like the detection of clothing and jewelry. The literature has also explored the analysis of other non-appearance-based features, extracting for example daily social network data [4]: information such as daily activities, logging emails, blogs, and handwriting can be used as features for classification. Such studies, however, are out of the scope of this work.

We propose a novel model-based approach for gender recognition that consists in extracting the parametric 3D human body model. The use of a model-based solution helps resolving the potential ambiguities that might arise when looking at aesthetic and appearance-based features only. In fact, the key goal of our work consists of using the SMPL[11] body-shapes parameters, which are invariant to clothing, hairstyle or other parameters commonly associated to one particular gender. In this way, we ensure that the model is sufficiently simple and reflects a standardized representation. In literature, only a few works address this problem using the body-shape information and, to the best of our knowledge, none of them use the parametric human body model SMPL[11]. In addition, most of the existing works use 3D human mesh vertices as features, significantly increasing the computational complexity, since the feature space that needs to be investigated is very large. In our case, the number of features is shrunk to only ten features. We also implement a CNN for comparison purposes, to evaluate our work against image-based methods. To do so, we use a pre-trained Resnet50[7] and we per-

form training and testing on the FVG[26] dataset, comparing the results against the ones obtained by the SVM, fed with the mesh parameters extracted from the video sequence.

The main technical contributions of this work can be summarized as follows:

- We propose an effective descriptor using the SMPL body shape parameter for gender recognition via a 3D human model. We prove that this type of classification is suitable for those datasets that are composed of 3D meshes, as well as videos, exhibiting the potential for the application in a wide set of use-cases, including video surveillance, robotics, and biometrics.
- We show how our classifier, with a reduced feature space, improves the results obtained by other model-based solutions proposed in the literature.

2 Related Work

Gender Recognition from body shapes. In literature, only a few works address the problem of gender classification using the body shape information. In fact, while 2D image data can be often misleading due to camera view point and image resolution, 3D shape models offer a more comprehensible description of the observed object (subject) at a negligible incremental cost. The authors in [21] propose a gender recognition solution based on 3D human body shapes obtained with laser scanning. The paper does not consider the full body, and the authors use multiple features extracted from the subjects' chest and torso. Furthermore, the authors assert in the conclusion that their approach fails in classifying overweight or fully dressed individuals. More recently, other works focus on the 3D mesh of the human body. The same authors present another research on gender classification in [22], where they perform the recognition task by considering the shape landmarks of 3D human body model. The work proposed in [25] considers the body shape as feature, and the classification relies on the geodesic distance on the mesh. They discover that the most relevant features are the geodesic distance between the chest and the wrist, as well as the one between the lower back and the face. The approach proposed in [16], introduces a 2D-vertex-based gender recognition model. The authors compare the performance of two classifiers, Support Vector Machines (SVMs) and Extremely Randomized Trees (ERTs). They obtain the most remarkable results by using as input feature the vertices of 3D mesh and the SVM as classifier, with an accuracy of 78%. Using a 3D vertex-based methods makes the feature space of the classifier very large. Originally, their meshes contained between 67290 and 68300 vertices; this required a re-sampling (using a uniform probability distribution), to the bottom side, namely 67290 vertices. Since this number was still very large to be processed, they extracted the most relevant features by using Principal Component Analysis (PCA), resulting overall in 350 components.

Gender Recognition from full-body images. In computer vision, gender recognition from whole-body images is a challenging task because the features extracted may not be discriminative enough for the classification and because background cluttering may be problematic. Gender classification has recently been addressed using convolutional neural networks. In [13], a CNN is trained considering the whole person body (Global CNN), the upper and then the lower portion of the human body (Local CNN). The Local CNN of the upper body achieves the highest accuracy because the face of a person

is more discriminative than the rest of the body. This is supported by a feature visualization method that shows where the CNN extracts the features on the image. When the face is not visible, features are concentrated in the rest of the body. In this case, the information is achieved from clothing, hairstyle, and body shapes information. Sometimes these features are not enough for accurate classification. This is confirmed also by Raza et al. in [18], where they propose an appearance gender recognition method where a deep neural network is used to extract the silhouette of the pedestrian image. The silhouette is then used as a binary mask to remove the background from the image. The outcome is fed into a stacked sparse autoencoder (SSAE). The gender is classified considering three different camera views (frontal, back, and mixed) and they obtain the lowest accuracy score, as expected, on the back view. The mixed views obtain an accuracy slightly lower than the one in the front. The frontal view is in fact more distinctive, as it contains information extracted not only from the body but also from the face. This proves that the body features may not be discriminative enough to reach the accuracy of face features. Ng et al. [13] show that by combining the Global and Local CNN from the upper part of the body it is possible to obtain a better model, outperforming the state-of-the-art methods.

Human Mesh Recovery from Natural Images. Model-based human pose estimation can be faced following two different approaches. Optimization-based methods iteratively fit a parametric human body model, e.g. SMPL [11], to estimate the body pose and shape of the 2D observations, usually 2D joints locations. This solution has been presented as an alternative to preexisting models coming from the scans of different bodies in a varied set of poses. With this model, Loper et al. [11] created realistic animated human bodies that represent different body shapes that deform naturally with pose and exhibit soft-tissue motions like those of real humans. In contrast, regression based methods use a deep network to directly estimate the model parameters from pixels. Both methods have some pros and cons. Optimization based methods tend to be very slow and sensitive to initialization. Regression based methods, instead of taking only a sparse set of 2D location, take into account all pixels values; at the same time, this leads to a mediocre image-model alignment, and a large quantity of data is usually necessary for training. Regarding the first approach, SMPLify [3] has been the first method that automatically estimates the 3D pose and shape of human body. The most recent works have focused on regression; in fact, since there is a deficiency of images with full 3D shape ground truth, alternative supervision signals to train the deep networks are searched. The majority of the solutions uses 2D annotations including 2D keypoints, silhouettes, or part segmentation. This information can be used as input [23], intermediate representation [14, 17], and supervision [8, 14, 17, 20, 23]. In this context, the SPIN algorithm [10], acronym of **SMPL oPtimization IN the loop**, presents a novel way of tackling the problem, finding a way to use the two methods in a collaborative fashion.

3 Datasets

Front View Gait Dataset (FVG). The FVG dataset [26] contains videos of 226 walking subjects, annotated by gender. It focuses only on the frontal view with three different

near frontal-view angles towards the camera and other variations in terms of walking, speed, carrying, clothing, cluttered background and time. The 226 subjects walk along a straight line of 16 meters toward the camera. The resolution of the video is full HD and the height of the person ranges from 101 to 909 pixels. For every subject, 12 videos have been captured, with different inclination of the camera (-45°, 0, 45°) and four variations of walking pace.

Archive of Motion Capture as Surface Shapes Dataset (AMASS). The AMASS dataset [12] consists of a collection of 15 MoCap datasets with gender annotation, represented with a common framework and parameterization. This has been achieved by converting the MoCap data into realistic 3D human meshes represented by a rigged SMPL body model, via the Mosh++ method.

Synthetic hUman foR REAL Dataset (SURREAL). The SURREAL dataset [24] contains 6 million frames of synthetic humans with ground truth pose, depth maps, segmentation masks, and gender information. The synthetic bodies are created using the SMPL body model. The SMPL parameters are fitted using the MoSh method from raw 3D MoCap marker data. The synthetic data has been generated rendering the following pieces of information: (i) a 3D human body model, whose pose was estimated with a motion capture system, (ii) a frame using background image, (iii) a texture map on the body, together with lightning and camera position. All these data are combined together in order to increase the diversity of the dataset.

4 Approach

The processing pipeline we propose consists of two stages: (i) extraction and preparation of the features, and (ii) classification. Since we are considering three different datasets, the pipeline slightly differs depending on which one is being used (see Fig.1). In particular, AMASS and SURREAL are characterized by parametric SMPL models; therefore the body shape parameters are given. For the FVG dataset, instead, an additional processing stage for features extraction is needed. This is performed by using the SPIN[10] algorithm, as follows:

- The parameters of the SMPL human parametric model are regressed with a deep network.
- These regressed values are used by an iterative fitting in order to align the model to the 2D keypoints.
- The fitted model is used as supervision for the network, closing the loop between the regression and optimization method.

The SMPL body model provides a function $\mathcal{M}\left(\vec{\beta},\vec{\theta}\right)$, that takes as input the body shape parameters $\vec{\beta}$ and the pose parameters $\vec{\theta}$, and gives as output the body mesh $M \in \mathbb{R}^{3N}$ with $N = 6890$ the number of vertices. The body pose is defined by a standard skeletal rig, composed by $K = 23$ joints; the pose is then defined by $|\theta| = 3 \times 23 + 3 = 72$ parameters (3 for each joint plus 3 for the root orientation). The body shapes of different people are represented by the function:

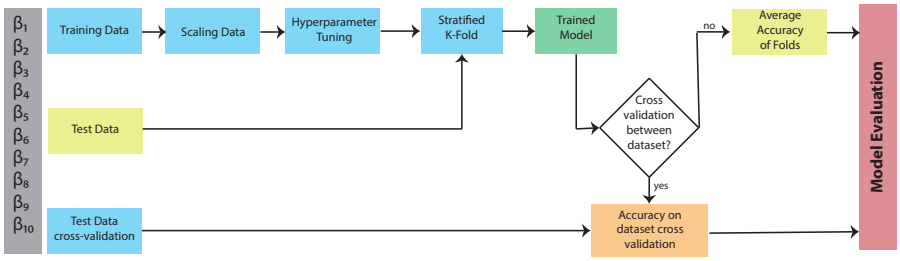

Fig. 2. Classification Pipeline. The training data are scaled and split in folds. On the training data the SVM Hyper-parameters are tuned. If the training and testing set belong to the same dataset, the accuracy of the model is the average accuracy over splits. Otherwise, the accuracy is calculated on the new testing set.

$$B_S\left(\vec{\beta};\mathcal{S}\right) = \sum_{n=1}^{|\vec{\beta}|} \beta_n \mathbf{S}_n \tag{1}$$

where $\vec{\beta} = [\beta_1, ..., \beta_{|\vec{\beta}|}]^T$, $|\vec{\beta}|$ is the number of linear shape coefficients, and the $\mathbf{S}_n \in \mathbb{R}^{3N}$ represents the orthonormal principal components of shape displacements. In the end, the body shape parameters are only ten and they can be defined as the principal components of the shape variation learned from 3D scans of thousands of people.

In summary, the main steps of the proposed methodology are listed hereafter:

1. The image is cropped, extracting the bounding box around the person using YOLO [19] as a detector. A bounding box is required by the SPIN algorithm, as it assumes that the person is centered in the image;
2. The cropped image is passed to the SPIN algorithm that extracts the body shape and pose coefficients;
3. The ten body shape coefficients are used as features for the classification, and split into training and test samples, following a cross-validation approach;
4. The training data is scaled and the tuning of hyper-parameters is performed;
5. Finally, the trained model accuracy is calculated.

4.1 Model Selection

In machine learning, we know that tuning the hyper-parameters is a key step, which allows building a robust and accurate model, preventing over/underfitting. In our implementation, we use the Grid Search method. We tune two grids: simple linear kernel, $\langle x, x' \rangle$ with five possible values of the regularization parameter, C and a RBF kernel with five different values of γ and four values of C. Since the chosen datasets exhibit a severe class imbalance we divide the data following a **stratified k-fold cross validation**; it consists of a variation of the k-fold method, where each fold is composed approximately by the same percentage of samples belonging to both classes. This allows us to mitigate the possible effects of gender classification, due to the gender unbalance.

The tuning of the parameters has been done in two steps. First, we calculate the most suitable number of splits dividing each dataset in a range of 3 to 10, and, for each of them, performing a model fitting. The final choice has been done by considering the number of splits that returns the highest accuracy. The second step consists in tuning the hyper-parameters and we choose the combination of parameters with the highest accuracy obtained from the confusion matrix. Finally, we proceed with model training over our different datasets with the hyper-parameters found. The accuracy of the model is calculated by averaging the accuracy of each split if the training and testing set belong to the same dataset, while with cross testing among different datasets the accuracy is calculated on the new testing set. The classification pipeline is illustrated in Fig.2. All the experiments have been conducted on a NVIDIA RTX 3090, using Pytorch for the network implementation and Scikit-Learn for the SVM implementation.

Table 1. Cross Validation Results. The experiments are conducted with different data splits: for example [FVG + A] - [A] means that the classifier has been trained on FVG and AMASS, and tested on AMASS. We also tested the algorithm adding progressively a larger amount of synthetic samples to AMASS and FVG. For example, [A + Sn] - [A]: n is the ID of the training set ($n = \{1, 2, 3, 4\}$); [A + Sn] -[A] means training on AMASS and SURREAL, and testing on AMASS. A larger ID number corresponds to a larger amount of SURREAL data.

Experiment	**#Train**	**#Test**	**#Female**	**#Male**	**Accuracy**(%)
[A]-[A]	317	159	68	91	84.23
[S]-[S]	3800	1900	977	923	99.94
[FVG]-[FVG]	5650	1130	415	715	87.38
[FVG + A] - [FVG]	214	79	22	57	83.5
[FVG + A] - [A]	214	104	58	46	95.2
[FVG + S1] - [FVG]	154	79	22	57	83.5
[FVG + S2] - [FVG]	183	79	22	57	86.1
[FVG + S3] - [FVG]	220	79	22	57	83.5
[FVG + S4] - [FVG]	294	79	22	57	84.8
[A + S1] - [A]	383	111	61	50	81.1
[A + S2] - [A]	455	111	61	50	82.9
[A + S3] - [A]	547	111	61	50	84.7
[A + S4] - [A]	730	111	61	50	86.5

5 Results

In this section, we describe the conducted experiments and the corresponding results, to validate the effectiveness of our classifier using the SMPL body shapes parameters for gender classification. We perform cross training and testing on three different types of dataset: synstetic, real and registration scans. In this way we want to demonstrate the effectiveness of the classifier on different type of data. We investigate the accuracy of the model when combining real and synthetic data from different datasets. The experimental results are listed in Table 1. The highest accuracy is obtained by the SURREAL

dataset, as we expected; SURREAL is a synthetic dataset and the body shape parameters have a perfect distribution between -5 and 5, making it a rather simple dataset to work with. As far as the FVG dataset is concerned, the returned accuracy is 87.38%. When we train and test on the AMASS dataset, the accuracy of classification decreases; so, even if this dataset is made of real registration scans, it consists of subject with a strong diversity in body shape. Instead the FVG dataset consists of real data, but its accuracy is higher than the AMASS dataset because it is characterized by subjects that do not strongly vary their body shapes. As far as the wrongly classified samples in AMASS is concerned, we assume that the performance decreases because the subjects are characterized by a sparse diversity in body shape.The failures in FVG occur generally when the subject is very far from the camera, namely exhibiting a reduced number of relevant pixels. This makes the extraction of the SMPL parameters with SPIN not sufficiently reliable. We then train and test the AMASS and FVG dataset adding in the training phase a progressively larger amount of synthetic data (from SURREAL): as we expected the accuracy increases when the synthetic data grow. As mentioned previously this is motivated by the fact that the synthetic samples are less subject to variations, making the classification easier and less prone to be adopted as substitutes for the real ones in this specific task.

5.1 Comparison with previous body shape-based methods

Since the novelty of our work consists in using the SMPL meshes, we could not find in the literature other works for a straightforward comparison. The available state-of-the-art papers [22, 25, 21] use the CAESAR dataset [2] characterized by meshes extracted through a laser scanner. We could not apply our method to these datasets because they are not characterized by SMPL mesh. Nevertheless, we still try to provide a fair comparison, although the differences between the meshes affect the features extracted for the classification. These features consist of the Geodesic Distances (GD) [25] between landmarks, which corresponds to the length of shortest path between two points constrained on the shapes, Normal Distributions (ND)[21] on the chest region, mesh Vertices Coordinates (VC) [16] and Landmarks Positions [22] (LP). Looking at

Table 2. Comparison with previous body shape-based methods. The term **RegS** stands for Registration Scans, **S** for Synthetic Shapes and **RealD** are Real Data (i.e. video data). **Dataset** indicates the train/test data, **Method** and **Features** the method and features used for the classification respectively. The results of our solution are listed in the last four rows.

Dataset	Type	Method	Features	Train	Test	Accuracy(%)	Pre-processing	Feature Space	Landmarks
CAESAR	RegS	[25]	GD	500	500	96.1	✗	11	✓
CAESAR	RegS	[21]	ND	1224	1224	96	✓	100	✗
CAESAR	RegS	[22]	LP	1224	1224	98.9	✗	219	✓
POSER[1]	RegS - S	[16]	VC	450	140	75	✗	350	✗
AMASS	RegS	**Ours**	SMPL	317	159	**84.23**	✗	10	✗
FVG	RealD	**Ours**	SMPL	5650	1130	**87.38**	✗	10	✗
AMASS - SURREAL	RegS-S	**Ours**	SMPL	5146	1030	**97.8**	✗	10	✗
FVG - SURREAL	RealD-S	**Ours**	SMPL	8987	1123	**92**	✗	10	✗

the methodology more in detail (see Table 2), the previous solutions require landmark detection or a pre-alignment process. They also have a much larger feature space. Instead, our method does not need any landmark or pre-processing step; furthermore, it has a much smaller feature space, resulting in a much faster computation. It is worth mentioning that our method can be effective also when using small datasets for training and testing, when compared to the ones used by the competing solutions. In addition, the proposed method uses a SMPL mesh that can also be extracted from a single image, thus it can be applied even when a laser scanner [25, 22, 21] is not available (e.g. surveillance), giving the solution generalization and scalability properties A fairer comparison can be conducted looking at the 3D vertex-based method presented in [16]. The authors achieve an accuracy of 75% using a very large number of features, even after feature reduction. With this respect, our method attains an accuracy of 87.38%, avoiding any feature reduction processes (e.g. PCA) since the SMPL mesh shrinks the feature space to only ten parameters.

5.2 Comparison with image-based methods

In order to prove the effectiveness of our solution, also when compared to image-based methods, we use a pre-trained Resnet50 and we train and test the architecture on the FVG dataset. The comparison is summarized in Table 3. The CNN reaches an average accuracy of 80% in the validation phase. When using the same dataset, our proposed method reaches 87%. This proves the peculiarity of the body-shape features used in this work with respect to the common features used by a simple CNN. This is also proved in [13], where the highest results is obtained when the face of the subject is visible (80.8%). In [18] they obtained an accuracy of 82.9% on frontal views and 82.4% on mixed views. In Fig.3 we can see examples of misclassified subjects by the CNN but correctly classified by our method considering the body shape parameters extracted from the 3D mesh. In fact, our solution does not rely on visual features and only considers the body shape information for gender classification, thus making it robust to camera pose changes, face appearance, and clothing. The last three columns report error in classification for both CNN and our method, possibly due to light conditions. For this reason, we made a use of a neutral body model for the incorrect classification samples only for visualization purpose, since the body model does not alter the values of body shape parameters.

6 Conclusions

We propose a novel approach for gender classification using SMPL body shapes parameters. This is suitable for all those datasets that are characterized by 3D meshes, as well as videos, exhibiting the potential for the application in a wide set of use-cases, as video surveillance, robotics, biometrics. Considering the low-dimensionality of the feature space that allows for fast computation, the proposed approach obtains satisfactory results, yet adding desirable properties, such as the use of a parametric mesh that provides a simple and a standard representation, with a number of vertices that is lower than the one used by competing methods. Our approach outperforms also the results of

Table 3. Comparison with image-based methods. We compare our proposed method against previous image-based works, as well as against the benchmark CNN we have implemented.

Method	**Dataset**	**View**	**Accuracy**(%)
Ng. et al.[13]	MIT[15]+APiS[27]	Upper frontal body Part	80.8
Ng. et al.[13]	MIT+APiS	Global + Upper frontal Parts	82.5
Raza et al.[18]	MIT+PETA[5]	Frontal	82.9
Raza et al.[18]	MIT+PETA[5]	Mixed	82.4
Our CNN	FVG	Frontal all body	80
Our Method	FVG	Frontal all body	**87.38**

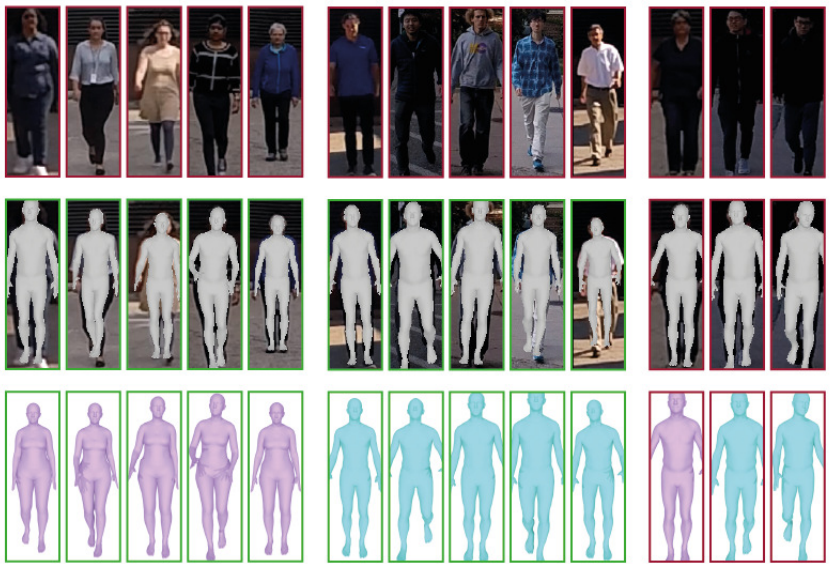

Fig. 3. Examples of classification output. The first row is characterized by subjects misclassified by the CNN. The second and third rows represent the classification output of our method. The red and green borders indicate respectively wrong and good classification output. The correct gender is indicated by the color of meshes in the third row.

image-based competing methods, since the features we adopt do not depend on camera view, and they are robust to face occlusion. In the future, our goal is to create a new dataset characterized by SMPL parametric shapes for gender recognition, as well as the recognition of additional attributes as, for example, age. We plan to use the DMPL [11] model, that has the same advantages of SMPL model but it considers the body deformations produced by the body movements and impact forces with the ground.

References

1. Poser - 3d character art and animation software - https://www.posersoftware.com/

2. Blackwell, S., Robinette, K., Boehmer, M., Fleming, S., Kelly, S., Brill, T., Hoeferlin, D., Burnsides, D.: Civilian american and european surface anthropometry resource (caesar). volume 2: Descriptions p. 192 (06 2002)
3. Bogo, F., Kanazawa, A., Lassner, C., Gehler, P., Romero, J., Black, M.J.: Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In: Computer Vision – ECCV 2016. Lecture Notes in Computer Science, Springer International Publishing (Oct 2016)
4. Burger, J., Henderson, J., Kim, G., Zarrella, G.: Discriminating gender on twitter. pp. 1301–1309 (01 2011)
5. Deng, Y., Luo, P., Loy, C.C., Tang, X.: Pedestrian attribute recognition at far distance. pp. 789–792 (11 2014). https://doi.org/10.1145/2647868.2654966
6. Dhomne, A., Kumar, R., Bhan, V.: Gender recognition through face using deep learning. Procedia Computer Science **132**, 2–10 (01 2018). https://doi.org/10.1016/j.procs.2018.05.053
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition (2015)
8. Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end recovery of human shape and pose (2018)
9. Kastaniotis, D., Theodorakopoulos, I., Economou, G., Fotopoulos, S.: Gait-based gender recognition using pose information for real time applications. In: 2013 18th International Conference on Digital Signal Processing (DSP). pp. 1–6 (2013). https://doi.org/10.1109/ICDSP.2013.6622766
10. Kolotouros, N., Pavlakos, G., Black, M.J., Daniilidis, K.: Learning to reconstruct 3d human pose and shape via model-fitting in the loop. Proceedings of the IEEE International Conference on Computer Vision pp. 1–10 (2019)
11. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: A skinned multi-person linear model. ACM Trans. Graphics (Proc. SIGGRAPH Asia) **34**(6), 248:1–248:16 (Oct 2015)
12. Mahmood, N., Ghorbani, N., Troje, N.F., Pons-Moll, G., Black, M.J.: AMASS: Archive of motion capture as surface shapes. In: International Conference on Computer Vision. pp. 5442–5451 (Oct 2019)
13. Ng, C.B., Tay, Y.H., Goi, B.M.: Pedestrian gender classification using combined global and local parts-based convolutional neural networks. Pattern Analysis and Applications **22** (11 2019). https://doi.org/10.1007/s10044-018-0725-0
14. Omran, M., Lassner, C., Pons-Moll, G., Gehler, P.V., Schiele, B.: Neural body fitting: Unifying deep learning and model-based human pose and shape estimation (2018)
15. Oren, M., Papageorgiou, C., Sinha, P., Osuna, E., Poggio, T.: Pedestrian detection using wavelet templates. pp. 193 – 199 (07 1997). https://doi.org/10.1109/CVPR.1997.609319
16. Pablo, N., Bruno, P., Celia, C., Virginia, R., Rolando, G.J., Claudio, D.: Gender recognition using 3d human body scans. In: 2018 IEEE Biennial Congress of Argentina (ARGENCON). pp. 1–6 (2018). https://doi.org/10.1109/ARGENCON.2018.8646293
17. Pavlakos, G., Zhu, L., Zhou, X., Daniilidis, K.: Learning to estimate 3d human pose and shape from a single color image (2018)
18. Raza, M., Sharif, M., Yasmin, M., Khan, M., Saba, T., Fernandes, S.: Appearance based pedestrians' gender recognition by employing stacked auto encoders in deep learning. Future Generation Computer Systems **88** (05 2018). https://doi.org/10.1016/j.future.2018.05.002
19. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection (2016)
20. Tan, V., Budvytis, I., Cipolla, R.: Indirect deep structured learning for 3d human body shape and pose prediction. In: BMVC (2017)
21. Tang, J., Liu, X., Cheng, H., Robinette, K.M.: Gender recognition using 3-d human body shapes. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) **41**(6), 898–908 (2011). https://doi.org/10.1109/TSMCC.2011.2104950

22. Tang, J., Liu, X., Cheng, H., Robinette, K.M.: Gender recognition with limited feature points from 3-d human body shapes. In: 2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC). pp. 2481–2484 (2012). https://doi.org/10.1109/ICSMC.2012.6378116
23. Tung, H.Y.F., Tung, H.W., Yumer, E., Fragkiadaki, K.: Self-supervised learning of motion capture (2017)
24. Varol, G., Romero, J., Martin, X., Mahmood, N., Black, M.J., Laptev, I., Schmid, C.: Learning from synthetic humans. In: CVPR (2017)
25. Wuhrer, S., Shu, C., Rioux, M.: Posture invariant gender classification for 3d human models. In: 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. pp. 33–38 (2009). https://doi.org/10.1109/CVPRW.2009.5204295
26. Zhang, Z., Tran, L., Yin, X., Atoum, Y., Wan, J., Wang, N., Liu, X.: Gait recognition via disentangled representation learning. In: In Proceeding of IEEE Computer Vision and Pattern Recognition. Long Beach, CA (June 2019)
27. Zhu, J., Liao, S., Lei, Z., Yi, D., Li, S.: Pedestrian attribute classification in surveillance: Database and evaluation. pp. 331–338 (12 2013). https://doi.org/10.1109/ICCVW.2013.51