Homologous Mesh Extraction via Monocular Systems

Mohamed Fateh Karoui[1,2(✉)] and Thorsten Kuebler[1]

[1] Human Solutions of North America, Morrisville, NC 27560, USA
Thorsten.Kuebler@human-solutions.com
[2] North Carolina State University, Raleigh, NC 27606, USA
mfkaroui@ncsu.edu

Abstract. Pose estimation of humanoid objects in monocular systems is a non-trivial problem that has been at the forefront of the human-computer interaction field. The ability for a computer to not only to detect the presence of a humanoid shape within an image but also to infer relative location and configuration has particular use for many applications. We explore a novel approach to solving this task by introducing a multi-stage preprocessing algorithm and a constrained pose estimator.

Keywords: Homologous · Mesh · Monocular · Convolutional neural network · Generative adversarial network · Regression

1 Introduction

Here at Human Solutions of North America, we have developed a novel multi-tiered approach to detecting and estimating poses in monocular system of humanoid objects using state of the art deep learning architectures and extensive domain knowledge through our commercial body scanners and Size North America proprietary data. Using a U-Net architecture we are able to segment an image to classify which pixels belong to a humanoid and which pixels belong to the background. The U-net architecture is ideal for this task and is considered the state of the art when it comes to image segmentation tasks. It is an encoder-decoder architecture that introduced a technique called a skip step that allows the propagation of feature locality throughout the network in order to classify what kind of subject a particular pixel belongs to. We then clip each detected subject and pass the image into a Convolutional Neural Network (CNN) to infer demographic information. This particular portion of the approach allows us to pick a good "initial guess" as to the structure of the subject. We extract information such as race, age, weight, and body morphology. Thusly, we choose a homologous mesh that has been statistically generated from our Size North America database for that particular demographic. The Size North America database consists of submillimeter precision three dimensional body scans of approximately 18,000 subjects distributed evenly across various demographics. This database allows us to produce a statistically representative three dimensional meshes of each demographic across multiple morphologies. Finally, we pass the homologous mesh into a deep neural network and

© Springer Nature Switzerland AG 2019
V. G. Duffy (Ed.): HCII 2019, LNCS 11581, pp. 182–197, 2019.
https://doi.org/10.1007/978-3-030-22216-1_14

produce a final mesh that represents the pose of the subject. This last step acts as a regressor and deforms the homologous mesh to fit the initial body pose of the subject.

This novel approach allows us to estimate the pose of multiple subjects that are within view of a monocular system as well as letting us infer a globally plausible body shape for occluded portions of the subject. This approach also opens the door for soft body simulation on subjects within an image. Applications of this methodology are wide and far impacting from three dimensional scene reconstruction and point of view visualization, to high fidelity motion capture from low cost systems.

2 Materials and Methods

2.1 Image Segmentation

Training a deep encoder-decoder neural network is rather tricky. This is caused by the conflicting nature of the requirements of the neural network versus the drawbacks of backpropagation. The U-Net architecture requires a maximization of information for semantic segmentation to be successful. This means that the standard methods of model regularization can no longer be utilized.

One major issue of deep neural networks is a tendency for overfitting. This is due to their large parameter space. The standard way to combat this issue is through dropout. During training we employ a process that stochastically stops gradients from propagating backwards through the layers in the neural network. This effectively kills neurons and forces the neural network to perform at a deficit. Many have theorized that this method causes the neural network to generate strong sub-classifiers in earlier layers. The late stage layers then ensemble these subnetworks to produce a final prediction. Unfortunately, structural information will be lost that act as input for later layer in the decoder network. Therefore, this method cannot be used.

To reduce computational cost, many neural networks, employ a Max Pooling layer whereby neurons of a previous layer are pooled together into a single neuron by taking the highest output signal from the group. This has the effect of reducing computational complexity while preserving the gross structure of the information. Unfortunately, local adjacency information is not preserved with this technique and fine image details that are important for classifying humans are lost.

We initially take a 512×256 three channel image, referred to as the source image, and pass it through a specialized "encoder-decoder" convolutional neural network referred to as a U-Net architecture [1]. The U-Net architecture introduces a tensor concatenation operator that allows structural information about identified classes to propagate throughout the neural network that is used to reconstruct a pixel-wise classification tensor. This concatenation operation is referred to as a "skip-step". Because we are dealing entirely with rank three tensors the concatenation operations occur along the third axis or the channels axis and are computationally cheap (Fig. 1).

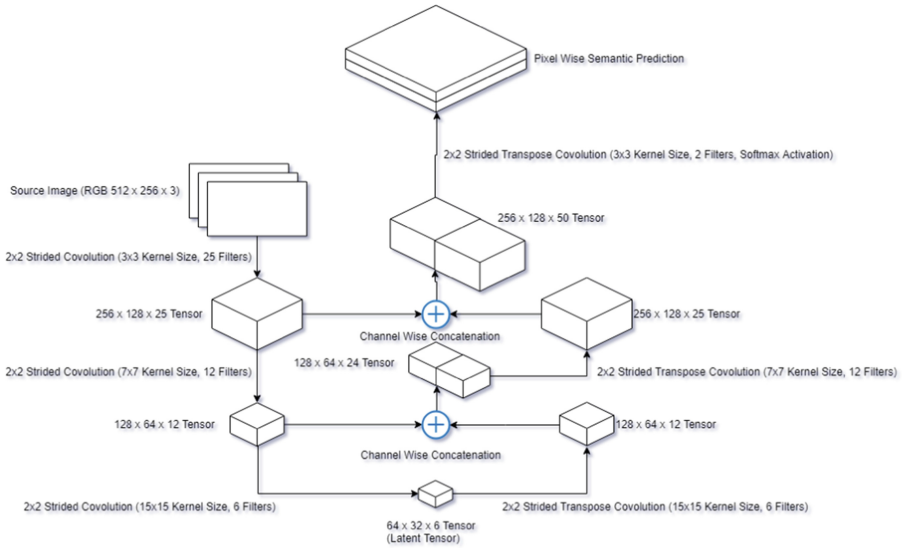

Fig. 1. Shows the general architecture for the U-Net Convolutional Neural Network. On the left hand side show the encoder network. On the bottom center is the latent tensor representation of the source image. On the right hand side is the decoder side of the neural network. In the center we have the concatenation operations that allow the structural information of the source image to propagate.

To reduce computational complexity, we employ a strided convolution that acts similarly to max pooling. The difference is our kernel size is always larger than the stride. This allows us to include adjacent information that is outside the "pooling" region while reducing the number of computations by power of two.

Since we are unable to use dropout to regularize our neural network we employed a method of streaming subsets of our original dataset, this is also referred to as incremental learning [1]. The Common Objects in Context (COCO) dataset [2], includes 330 thousand images that are semantically labeled by object class. The dataset is excellently curated and provides a large variety of examples to train on (Fig. 2).

Our activation function, which provide the non-linear capacity for our neural network, was chosen specifically to remove the need for batch normalization [3]. SELU, or scaled exponential linear units belong to a class of self-normalizing activation functions. This activation function allowed us to remove the need for additional normalization layers without losing the benefit that normalization has to solving the vanishing gradient property.

SELU is defined as

$$f(x,\alpha) = \lambda \begin{cases} \alpha(e^x - 1), & x < 0 \\ x, & x \geq 0 \end{cases}$$

Where λ is a learned parameter that acts as a scaling factor to boost gradient propagation.

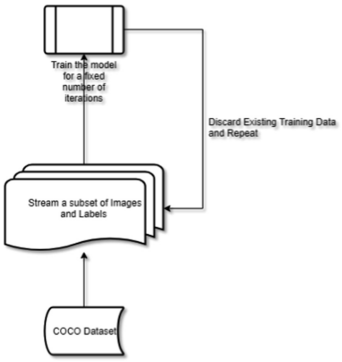

Fig. 2. The iterative training process allows us to define a dynamic set of images to train on. This removes the issue of training over fit without having to perform dropout and other model regularization techniques.

2.2 Clipping

Once a class is identified within the source image we must clip the class object into a separate image to extract demographic information. This clipped form of the image isolates the subject from external sources of information that may add undue noise during the subsequent processes.

Clipping is performed using a masking methodology on a low-pass canny filter. Initially we take a source image and pass a Gaussian Kernel Convolution across the source image to remove high frequency information from the image. This will have the effect of reducing the number of possible edges, as shown in Fig. 3.

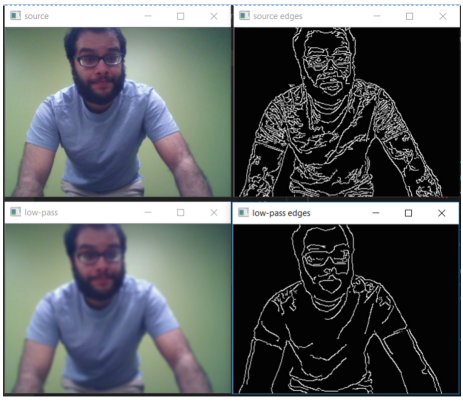

Fig. 3. Shows the canny edge filter as applied directly on the source image (top right) versus being applied after a low pass filter operation on the source image (bottom right).

Once we extract edges we apply a pixel-wise multiplication of our region proposal. The result of the operation yields a very clean image that contains only the subject to be passed on later processes (Fig. 4).

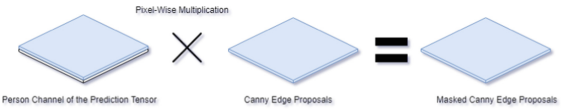

Fig. 4. Edge masking allows us to focus on edges that we think belong to a human.

2.3 Demographics Estimation

The demographics of a detected subject plays an important role in selecting the right initial conditions for the mesh regression procedure. Extracting the demographics of a subject is done using three convolutional neural networks. Each one is responsible for extracting a prediction for age, race, and gender. The CNN's use two principles to achieve better than human performance when classifying demographics. A decaying special drop rate, and an expanding kernel size.

To regularize the neural networks and prevent over fit, we employ a high drop rate in the earlier stages of the neural network and a low dropout rate in the later stages of the neural network. This improves the ability of strong subnetworks to be generated for extracting low level features. In the later stages we want the layers to act as an ensembling mechanism. Secondly, expand the kernel sizes to capture local features within the image at earlier stages and global features in later stages.

The result of the convolutional neural networks is then concatenated to produce a final prediction vector to be used in subsequent steps (Fig. 5).

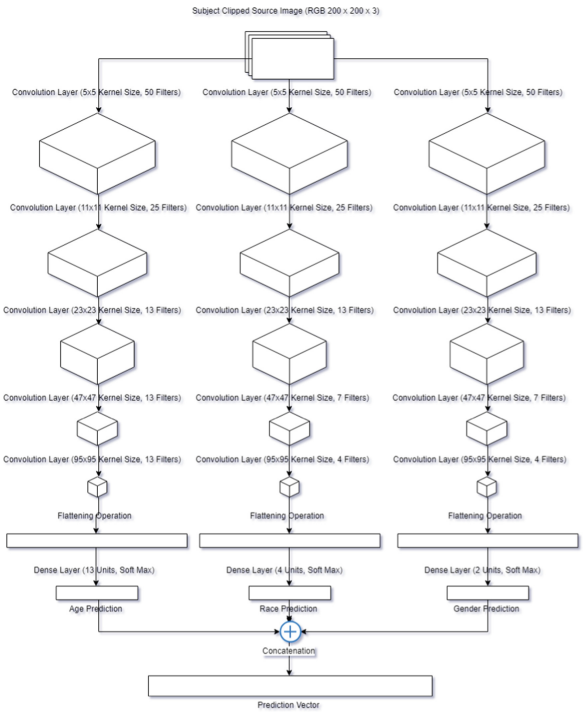

Fig. 5. Highlights the key architecture of the set of Convolutional Neural Networks that are responsible for extracting demographic information from the subject after clipping.

2.4 Homologous Mesh Generation

During our product developments we conducted a size survey called Size North America which consisted of scanning eighteen thousand diverse subjects using millimeter precision body scanners. The subject takes a quick demographic survey and then change into skin-tight under garments. They then enter our body scanner whereby multi-laser optical measurements occur across the entire length of the body producing High Density Point Cloud (HDPC) data. Using propriety software, we aggregated our HDPC data into statistically representative and vertex uniform meshes called homologous meshes (Fig. 6).

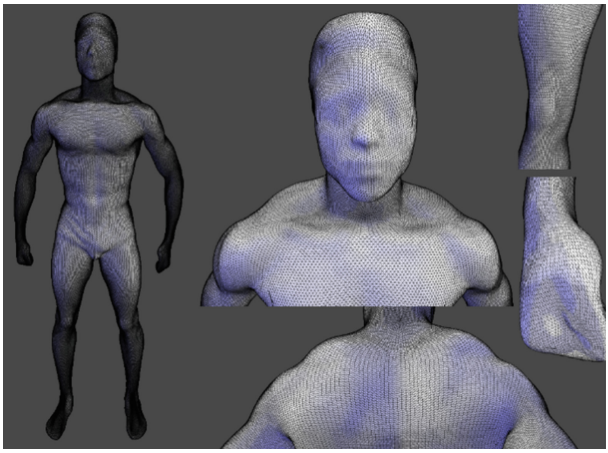

Fig. 6. Showcases the vertex uniformity of the homologous meshes within our dataset.

2.5 Homologous Mesh Estimation

Given a demographic prediction vector P_i about a particular subject then a reasonable estimate about a subject's mesh M_i can be given by an inner product of the prediction vector with the basis B of the space representing all possible human meshes. We approximate the basis of this space using out homologous mesh's extracted from our Size North America survey.

$$M_i = \frac{P_i \cdot B}{P_i \cdot P_i}$$

Where B is the basis set of meshes defined as

$$\left\{B_{g,r,a} | B_{g,r,a} \in \mathrm{M}^{\mathrm{n\,x\,3\,x\,160,785,}}, g \in \mathrm{Z}\ r \in \mathrm{Z}, a \in \mathrm{Z}\right\}$$

and P_i is the prediction vector defined as

$$\{P_i | P_i \in \mathrm{R}^n\},$$

for a subject i (Fig. 7).

Fig. 7. Shows a sample of our homologous meshes across demographic range. Starting from the top we show meshes for Female African Americans, Male African Americans, Female Asians, Male Asians, Female Others, Male Others, Female White, Male White. Each mesh across a row is a statistically representative model of our age group classes. Starting from the left we show meshes for ages 0–11, 12–17, 18–23, 24–29, 30–35, 36–41, 42–47, 48–53, 54–59, 60–65, 66–71, 72+ respectively.

In essence this process is a weighted average operation of all the homologous meshes across our demographic classes. The weights are determined by the probabilities produced by the neural network.

2.6 Pose Estimation

Pose estimation was accomplished using a Convolutional Neural Network on clipped source images. Preprocessing the image to remove background information allowed us to reduce the complexity of our neural network. Since pre-clipping removes background information, our neural network did not need to learn what a person looks like.

We posit that the pose estimator works by simply regressing a central skeleton into the contour provided. Our neural network's final layer simply had 22 degrees of

freedom. We constructed a constrained skeleton layer based on pre-existing anatomical models which greatly reduced the regression times and improved overall accuracy when compared to a standard dense layer output. Our constraints are defined by medically accepted normal ranges of motion (Tables 1, 2, 3, 4, 5, 6 and 7).

Table 1. Describes the normal range of motion for the hip.

Type	Minimum angle (Deg.)	Maximum angle (Deg.)
Flexion	0	125
Extension	115	0
Hyperextension	0	15
Abduction	0	45
Adduction	45	0
Lateral rotation	0	45
Medial rotation	0	45

Table 2. Describes the normal range of motion for the knee.

Type	Minimum angle (Deg.)	Maximum angle (Deg.)
Flexion	0	130
Extension	120	0

Table 3. Describes the normal range of motion for the ankle.

Type	Minimum angle (Deg.)	Maximum angle (Deg.)
Plantar flexion	0	50
Dorsiflexion	0	20

Table 4. Describes the normal range of motion for the foot.

Type	Minimum angle (Deg.)	Maximum angle (Deg.)
Inversion	0	35
Eversion	0	25

Table 5. Describes the normal range of motion for the shoulder.

Type	Minimum angle (Deg.)	Maximum angle (Deg.)
Flexion	0	180
Extension	0	50
Abduction	0	90
Adduction	90	0
Lateral rotation	0	90
Medial rotation	0	90

Table 6. Describes the normal range of motion for the elbow.

Type	Minimum angle (Deg.)	Maximum angle (Deg.)
Flexion	0	160
Extension	145	0
Pronation	0	90
Supination	0	90

Table 7. Describes the normal range of motion for the wrist.

Type	Minimum angle (Deg.)	Maximum angle (Deg.)
Flexion	0	90
Extension	0	70
Abduction	0	25
Adduction	0	65

2.7 Rigging Homologous Meshes

Once pose estimation is complete applying the pose to the mesh involves regressing the mesh skeleton which applies a system of linear transformations to the mesh allowing the mesh to be regressed into the desired pose.

To simplify the rigging process of the homologous mesh we used the software Unity. By defining the key points of a skeleton we are able to apply transformations to the entire mesh through the 3D rending software (Fig. 8).

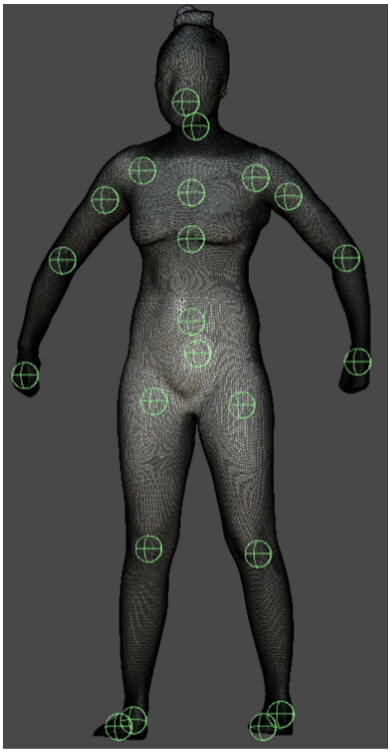

Fig. 8. Showcases the control points of the skeleton defined in Unity. These will act to define a system of linear transformations that will be applied to each vertex on the mesh.

3 Results

The image segmentation network was particularly difficult to train as great care had to be taken when dealing with class weights. Code was developed to dynamically calculate class weight upon each batch. The class weights were calculated by counting pixels belonging to people versus pixels belonging to the background. This added procedure cause training times to be much higher, but yielded very good results (Fig. 9).

The clipping operation yielded expected results whereby 83% of human subjects in validation data were clipped from the source image. This is largely sufficient for images in the wild. We expect the use case for this algorithm to be mostly situated in controlled well lit environments (Fig. 10).

Fig. 9. Showcases very hard validation examples of the image segmentation process. Input images are shown in column 1, the ground truth labels in column 2, and the neural network results in column 3. Background pixels are represented in green, while pixels belonging to people are represented in blue. (Color figure online)

Fig. 10. Shows a sample of the clipping process in a non-trivial test case where the subject has intersecting edges with a background. The subject also has a wide variety of occluding features such as facial hair with no discernable variation from his shirt.

Training the demographic convolutional neural networks yielded a significantly greater than random accuracy for each network. We trained these networks using the UTKFace dataset [16] which provides a well curated set of faces with race age and gender annotations (Figs. 11, 12 and 13).

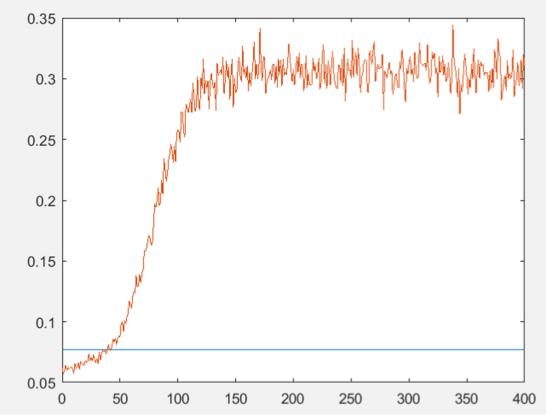

Fig. 11. Outlines the train curve for extracting age estimations after 400 training epochs. The top one prediction accuracy for 13 classes plateaued after the 150th epoch. The jitter in accuracy is caused by the dropout rate in earlier layers as compared to the training step size (1e-3). The line in blue shows the benchmark accuracy if the neural network were to classify age at random. (Color figure online)

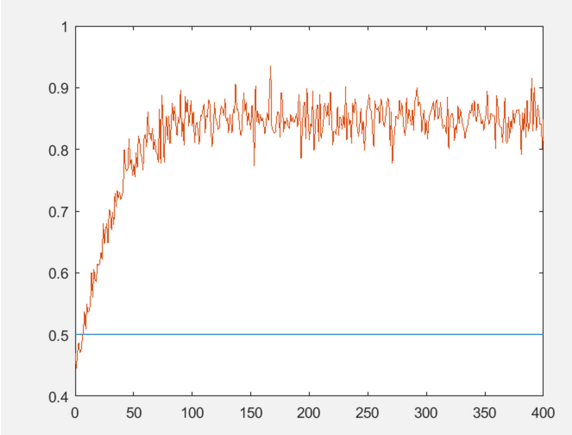

Fig. 12. Outlines the train curve for extracting gender estimations after 400 training epochs. The top one prediction accuracy for two classes plateaued after the 60th epoch. The line in blue shows the benchmark accuracy if the neural network were to classify gender at random. (Color figure online)

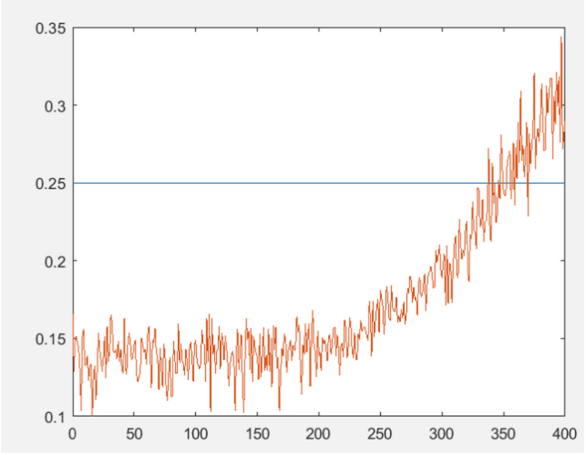

Fig. 13. Outlines the train curve for extracting race estimations after 400 training epochs. The top one prediction accuracy for four classes did not plateau and had significant trouble attaining greater than random accuracy. The line in blue shows the benchmark accuracy if the neural network were to classify race at random. (Color figure online)

4 Discussion

4.1 Improvements

One major drawback that this methodology has is the multistage approach. Computationally speaking this is not efficient and may suffer when implemented on lower end hardware. We propose that the whole process be integrated into a single feed forward neural network.

Our neural network size was also limited by the capabilities of our hardware. Source images were down sampled from their original sizes. Therefore, it is reasonable to expect a major loss of fine details that are crucial to the process. Expanding the number of filters and adding more layers may allow the neural network to perform better.

4.2 Applications

Mobile Sizing

Our methodology opens the door for robust sizing estimations of a subject without the need for expensive hardware. Given a proper reference point the algorithm can extract length measurements across any set of points defined along the mesh. This has direct applications for the fashion, automotive, aerospace, and ergonomic industries.

An example scenario for the fashion industry would be at the retail level. A boutique fashion store can setup or use existing camera systems to build mesh estimates for all their customers. When a customer selects a garment they can instantly view a simulation of how the garment looks and moves on their body. This removes the risk of

exposing expensive apparel to the customer and allows the store to reduce inventory while catering to a higher range demographics.

In ergonomic research, a key area of the field that is lacking is the ability to rapidly prototype designs on computer systems. The ability to develop ergonomic products that can be easily be tested on specific demographics plays an important role. Our technology has particular use when the designer(s) does not have access to an expensive demographics sizing database. They will be able to easily produce simulation ready meshes from any images.

Social Networking and Information Pivoting
The ability search information broker databases allow one to leverage limited knowledge about a subject to expand their information. Unfortunately, traversing these databases becomes an intractable computational nightmare. Searching social media databases is nearly impossible when looking for a particular subject. The ability to narrow down the search space for a human subject greatly reduces search times.

When this methodology is paired with other information gathering techniques, such as natural language processing, one may be able to extract knowledge about a human subject just by having a simple conversation with the subject. This has direct applications in law enforcement. During an interrogation the interviewer's task is to extract information that might otherwise be hidden or obscured. Real time information validation plays a very crucial role. Our system can be used to search and validate a person's identity in real time. Information such as age, gender, race, height, body morphology can be used as filtering terms to search offender databases without the need to rely on the human subject to provide accurate information.

Motion Capture
The motion capture industry has barrier of entry in terms of cost of equipment and education. High fidelity motion capture systems requires dedicated studios with dedicated hardware and a very knowledgeable team to maintain [18]. With our pose estimation and mesh regression we are able to produce reasonably accurate motion capture that can later be fed game development projects and movie studios. Our system's ability to produce homologous mesh's allows for easy integration with pre-existing animation and rendering pipelines. The vertex uniformity of the mesh lets studios perform soft-body and hard body simulations to produce highly realistic scene renderings at a fraction of the cost.

4.3 Privacy Implications

The sensitive nature of extracting demographic data from images has grave privacy implications. The applications for this technology should be selected to align with the public good. Such a technology could be used to leverage into personal and private details. The methods discussed by this paper are not the edge cases for the potential application of this technology. Such methods can be used to estimate data protected by legislation such as medical history. With the right combination of inputs bad actors may use this technology to perform identity theft and other more malicious acts.

Age has particularly strong privacy implications if this technology is used in public facing systems. The ability to extract identifying features from the minority subset of

the population without parental approval can breach many local and federal regulations. Such a system must have filters in place to ignore subjects that have reasonable evidence that they are below the age of majority.

Race plays an important role in the system's ability to extract fine details with a high degree of accuracy. Initial structural features that reduce regression times are highly dependent on race. There are many downsides to a system that relies on accurately classifying race. If the convolutional neural network is trained on data that has a class imbalance between races the network may miss-identify a race or the race in particular may become under or over represented within the prediction vector. This will negatively impact the quality of the results. In terms of morality, threat analysis systems and the like that rely on race for identification and classification may compound race inequalities. Therefor the author proposes that systems that are used to predict human behavior should abstain from using race qualifiers.

Gender, like race, is a predictive qualifier for estimating body structure. The very trivial example is bust size. If an initial guess for a female subject was not statistically representative for a female, the regressor would likely need more iterations for a fixed step size to optimize the initial mesh to fit a female bust. Choosing the correct gender is crucial to an accurate representation of a subject. Unfortunately, it is very difficult to represent the subset of the population that is gender ambiguous. By the very definition a transgender subject crosses the boundaries between classes and can cause even the most perceptive humans to think twice. This poses a very difficult technological problem and may also exacerbate the political issues around transgender rights.

Many of the examples presented show the need for a good demographic classifier, but we must take particular care when these systems are applied to public applications. We must not give public institutions and regulatory bodies technological justifications to widen the gap of inequality. Nor must we employ these technologies prematurely when they have a direct impact on a person's life and liberty.

Acknowledgements. A large effort was made to collect the data presented here in the paper. The authors would like to thank Human Solutions of North America for the continuous effort to provide high quality scan data. In particular Dylan Hendricks, a key account manager at Human Solutions of North America, for his critical role in the generation of the homologous meshes. The authors would also like to extend great appreciation for the open source community that has developed around the field of machine learning. Their continued effort in providing key systems made our research possible.

References

1. Pasquale, G., Ciliberto, C., Odone, F., Rosasco, L., Natale, L.: Are we done with object recognition? The iCub robot's perspective. Robot. Auton. Syst. **112**, 260–281 (2019)
2. Lin, T.Y., et al.: Microsoft COCO: common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 740–755. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10602-1_48
3. Ruiz-Garcia, A., Elshaw, M., Altahhan, A., Palade, V.: A hybrid deep learning neural approach for emotion recognition from facial expressions for socially assistive robots. Neural Comput. Appl. **29**(7), 359–373 (2018)

4. Lee, D., Nakamura, Y.: Motion recognition and recovery from occluded monocular observations. Robot. Auton. Syst. **62**(6), 818–832 (2014)
5. Sarafianos, N., Boteanu, B., Ionescu, B., Kakadiaris, I.A.: 3D Human pose estimation: a review of the literature and analysis of covariates. Comput. Vis. Image Underst. **152**, 1–20 (2016)
6. Kang, M.-J., Lee, J.-K., Kang, J.-W.: Combining random forest with multi-block local binary pattern feature selection for multiclass head pose estimation. PLoS ONE, 1–24 (2017). https://doi.org/10.1371/journal.pone.0166749
7. Ariz, M., Villanueva, A., Cabeza, R.: Robust and accurate 2D-tracking-based 3D positioning method: application to head pose estimation. Comput. Vis. Image (2019, in press)
8. Zhu, S., Mok, P.Y., Kwok, Y.L.: An efficient human model customization method based on orthogonal-view monocular photos. Comput. Aided Des. **45**(11), 1314–1332 (2013)
9. Yu, J., Guo, Y., Tao, D., Wan, J.: Human pose recovery by supervised spectral embedding. Neurocomputing **166**, 301–308 (2015)
10. Kim, H., Lee, S.-H., Sohn, M.-K., Kim, D.-J.: Illumination invariant head pose estimation using random forests classifier and binary pattern run length matrix. Hum.-Centric Comput. Inf. Sci. **4**(1), 9 (2014)
11. Zhang, Z., Zhao, R., Liu, E., Yan, K., Ma, Y.: Scale estimation and correction of the monocular simultaneous localization and mapping (SLAM) based on fusion of 1D laser range finder and vision data. Sensors **18**(6), 1948 (2018)
12. Piccirilli, M., Doretto, G., Adjeroh, D.: Framework for analyzing the whole body surface area from a single view. PLoS ONE, 1–31 (2017). https://doi.org/10.1371/journal.pone.0166749
13. Droniou, A., Ivaldi, S., Sigaud, O.: Deep unsupervised network for multimodal perception, representation and classification. Robot. Auton. Syst. **71**, 83–98 (2015)
14. Hu, R., Savva, M., van Kaick, O.: Functionality representations and applications for shape analysis. Comput. Graph. Forum **37**(2), 603–624 (2018)
15. Hussein, A., Elyan, E., Gaber, M.M., Jayne, C.: Deep imitation learning for 3D navigation tasks. Neural Comput. Appl. **29**, 389–404 (2018)
16. Zhang, Z.S.: Age progression/regression by conditional adversarial autoencoder. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
17. Jáuregui, D.A.G., Horain, P.: Real-time 3D motion capture by monocular vision and virtual rendering. Mach. Vis. Appl. **28**(8), 839–858 (2017)
18. Kim, Y., Kim, D.: Real-time dance evaluation by markerless human pose estimation. Multimedia Tools Appl. **77**(23), 31199–31220 (2018)
19. Basu, S., Poulin, J., Acton, S.T.: Manifolds for pose tracking from monocular video. J. Electron. Imaging **24**(2), 023014-1–023014-21 (2015)
20. Biasotti, S., Cerri, A., Bronstein, A., Bronstein, M.: Recent trends, applications, and perspectives in 3D shape similarity assessment. Comput. Graph. Forum **35**(6), 87–119 (2016)
21. Calderita, V.L., Bandera, J.P., Bustos, P., Skiadopoulos, A.: Model-based reinforcement of kinect depth data for human motion capture applications. Sensors **13**, 8835–8855 (2013)
22. Chen, C., Zhuang, Y., Xiao, J.: Silhouette representation and matching for 3D pose discrimination – a comparative study. Image Vis. Comput. **28**(4), 654–667 (2010)
23. Xia, S., Gao, L., Lai, Y.K., Yuan, M.-Z., Chai, J.: A survey on human performance capture and animation. J. Comput. Sci. Technol. **32**(3), 536–554 (2017)
24. Xu, C., Nanjappa, A., Zhang, X., Cheng, L.: Estimate hand poses efficiently from single depth images. Int. J. Comput. Vis. **116**, 21–45 (2016)