2018

Machine Learning Approaches to Human Body Shape Analysis

Marco Piccirilli

MACHINE LEARNING APPROACHES TO HUMAN BODY SHAPE ANALYSIS

Marco Piccirilli

Thesis submitted to the
Benjamin M. Statler College of Engineering and Mineral Resources
at West Virginia University

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Electrical Engineering

Donald Adjeroh, Ph.D., Committee Chairperson
Gianfranco Doretto, Ph.D., Co-Chair
Arun Ross, Ph.D.
Bojan Cukic, Ph.D.
Natalia Schmid, Ph.D.
Peter Giacobbi, Ph.D.
Xin Li, Ph.D.

Lane Department of Computer Science and Electrical Engineering

Morgantown, West Virginia
2018

Keywords:Computer Vision, Machine Learning, Computational Geometry, Soft Biometrics, Human Body

ABSTRACT

Machine Learning Approaches to Human Body Shape Analysis

Marco Piccirilli

Soft biometrics, biomedical sciences, and many other fields of study pay particular attention to the study of the geometric description of the human body, and its variations. Although multiple contributions, the interest is particularly high given the non-rigid nature of the human body, capable of assuming different poses, and numerous shapes due to variable body composition. Unfortunately, a well-known costly requirement in data-driven machine learning, and particularly in human-based analysis, is the availability of data, in the form of geometric information (body measurements) with related vision information (natural images, 3D mesh, etc.). We introduce a computer graphics framework able to generate thousands of synthetic human body meshes, representing a population of individuals with stratified information: gender, Body Fat Percentage (BFP), anthropometric measurements, and pose. This contribution permits an extensive analysis of different bodies in different poses, avoiding the demanding, and expensive acquisition process. We design a virtual environment able to take advantage of the generated bodies, to infer the body surface area (BSA) from a single view. The framework permits to simulate the acquisition process of newly introduced RGB-D devices disentangling different noise components (sensor noise, optical distortion, body part occlusions). Common geometric descriptors in soft biometric, as well as in biomedical sciences, are based on body measurements. Unfortunately, as we prove, these descriptors are **not** pose invariant, constraining the usability in controlled scenarios. We introduce a differential geometry approach assuming body pose variations as isometric transformations of the body surface, and body composition changes covariant to the body surface area. This setting permits the use of the Laplace-Beltrami operator on the 2D body manifold, describing the body with a compact, efficient, and pose invariant representation. We design a neural network architecture able to infer important body semantics from spectral descriptors, closing the gap between abstract spectral features, and traditional measurement-based indices. Studying the manifold of body shapes, we propose an innovative generative adversarial model able to learn the body shapes. The method permits to generate new bodies with unseen geometries as a walk on the latent space, constituting a significant advantage over traditional generative methods.

“Studere studere....., post mortem quid valere?”

(cit. Mautilio)

“Memento Audere Semper!”

(cit. D’Annunzio)

Acknowledgments

I would like to express my immense gratitude to my parents for their continuous support, and the motivation to look always ahead, with passion and determination. My sincere thanks go to my advisors for their patience and the long years of research. Finally, I thank my fellow labmates for the long discussions and the great camaraderie.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Human Body Shape Analysis: A Vision-Driven Approach

Measuring the body of humans is a significant activity, which has long been performed to describe subjects for a variety of very different tasks. In biometrics, for instance, we are interested in finding some stable descriptors to identify, verify, or classify subjects. Nutritionists and physicians are interested in body indexes capable of assessing a patient's health status. Ergonomists and stylists are interested in the body dimensions to design accessories, equipment, clothes, and comfortable spaces. To all these fields, the **compact** and **robust** representation of the body shape is fundamental. Traditional techniques have been used for years to describe the body shape and are still in use in many fields of everyday life. However, despite the introduction of many useful solutions, a compact description is always a hard problem due to the large number of poses that a body can assume, the vast variety of shapes, and nonetheless, the perceived appearance of the body from different views. A successful and reliable solution for this highly nonlinear problem will constitute the holy grail, not just for one discipline, but for a good section of the modern society, where style and appearance are at the center of our lives.

The human vision system can process (visual) signals in a fraction of the time that will be required by other systems (smell, hearing, taste, touch). This formidable capability places visual information at the center of all human activities, but it also makes understanding of the visual system tremendously complex.

Due to the still increasing importance of the visual information, the field of computer vision is having an exponential growth in research and publication, boosted by the recent technological advances that permit storage and signal processing unthinkable a decade ago. Machine learning, moreover, is another field taking advantage of the technological advance that is contributing to the computer vision boost. The refreshed **deep** techniques are playing a major role in the success of computer vision techniques.

1.2 Human Body Shape Analysis: A Soft-Biometrics Viewpoint

Soft Biometrics [169] is defined as any anatomical or behavioral characteristic that provides some information about the identity of a person, but that is not sufficient to identify the subject. Gender, ethnicity, age, height, weight, eye color, scars, marks, tattoos, and voice accents are typical **soft** biometrics traits. Typically, **soft** biometrics is often used as a complement to traditional **hard** biometrics (fingerprint, iris, face, etc.) to improve the recognition accuracy. More recently, soft biometrics has had a life on his own with the advent of surveillance systems and long-range cameras (NIR, LF IR) where the traditional biometric traits are not available, and due to the uncooperative nature of the acquisition, only soft biometric traits are useful.

Soft biometrics, however, present different problems concerning reliability and accuracy. Combining many traits (gender, ethnicity, age, height, etc.), soft biometric systems usually lack persistence: the anthropometric features (e.g., height) can vary significantly for the same age

group (intra-group variation). They also lack distinctiveness: skin color or eye color cannot be used for distinguishing between individuals with the same ethnicity (inter-group variation). Finally, the considerable time, effort, and training required to get reliable measurements is a major cause of errors in the measurements. Two important challenges need to be addressed to effectively incorporate the soft biometric information into the traditional biometric framework. The first challenge is the automatic and reliable extraction of soft biometric information in a nonintrusive manner, without causing any inconvenience to the users, which we'll study in this thesis. The second, the fusion with primary biometrics, is out of the scope of this work.

Anthropometric soft biometric systems have been shown to obtain good results. In [3] we have shown that soft biometrics system can be used successfully in challenging situations. In particular, we have assessed the correlation and predictability of body measurements in a population of individuals. Using three seed measurements to predict the other 41 measurements, and using both measurements for gender prediction produced a classification rate of 88.9 % on the testing set.

Although we obtained encouraging results on the CEASAR [244] and MoCap [67] datasets, we had less reliable results on our small acquisition and contradictory performances on the data from video. For the first dataset, we attribute the poor performance to different population age. Our acquisition, composed mostly of undergraduate and graduate students was quite different from the CAESAR dataset [244], comprised mostly of adults. More interesting were the results using the measurements from video. We quickly realized that traditional body measurements, using devices like Microsoft Kinect [285] were less stable, with higher relative error than hand-made measurements. In the next section, we'll review some essential works on soft biometrics and present the motivation for our work.

1.2.1 Related Work in Soft-Biometrics

The first biometric system built by Alphonse Bertillon in 1883 [135] used anthropometric features, such as the length and breadth of the head and the ear, length of the middle finger and foot, height, along with attributes like eye color, scars, and tattoo marks. These measurements were obtained manually. Although (intra-user) variability was observed, a combination of several measurements was sufficient to identify a person with reasonable accuracy. This biometric system can be considered as the first soft biometrics system by modern definitions, later was replaced by a fingerprint-based system [135].

Recently, there has been an increased interest in soft biometric features, though the robust extraction of these features is still an open problem. When traditional biometrics features are available, soft biometric traits can be extracted more efficiently. For instance, given the face image, various attributes can be extracted with sufficient reliability, e.g., gender [50], ethnicity [111], age [154, 156], and eye color. However, the need for the primary biometric features is a key limitation. Soft biometric systems are reviewed in recent surveys by Dantcheva et al. [77, 78], Nixon et al. [208, 239], and others [131, 243].

Between the anthropometric measurements, the stature is the easiest to acquire. However, depending on the acquisition device, different challenges are encountered. Criminisi, taking advantage of the well-known work on single view metrology [72], developed an uncalibrated method for stature measurement [73]. Nguyen et al. [207] used a new technique called cross-ratio in parallel with the vanishing point method, for static stature measurement, and dynamic measurement when the subject is walking. Another crucial area of soft biometrics is weight prediction. Cao et al. [49] predicted weight and gender using a copula model with measurements taken from the CAESAR dataset [244]. Velardo et al. [290], inspired by [264] on height estimation, proposed a model-based approach to correlate the weight with common anthropometric measurements. Unfortunately, the analysis was based on hand-made measurements from

the NHANES [56] dataset, and on a limited set of RGB images. The anthropometric measurements from the images were extracted manually assuming an oval shape for the body section. Although the method produces good results, the approach is far from being automatic. In [291] the same authors extended the former method using a neural network approach, instead of a multilinear regressor. The datasets used are the well known NHANES [56] dataset, and a new acquisition with the Microsoft Kinect RGB-D sensor [285]. This new dataset, however, was limited in size, to only 15 subjects. The method shows the sound capabilities of the Kinect sensor, but the small RGB-D dataset limits the evaluation of the results to a restricted number of body shapes.

Recently Madadi et al. [189] presented a novel method to extract anthropometric measurements using depth sensors, and the body parts tracking algorithm [266]. This method assumes a multi-parts labeled training dataset, and that the subject is aligned to the best model in the dataset. These constraints, although familiar to many 3D matching frameworks, make this approach quite limited, and not scalable to a high number of poses.

1.3 Human Body Shape Analysis: A Medical Science Viewpoint

In medical sciences, the human body is at the center of all analysis. From the ancient Greek culture to Leonardo' Vitruvian man, and to the modern age, the human body composition, functions, and shape have been deeply studied. A portion of today' studies regards the understanding of, and the fight against, important diseases. Today, a significant focus is on the role of body shape in the understanding, prediction, and fight against important diseases.

In the last decades, medicine and biomedical focus was on developing efficient and vital diagnostic tools for use by every physician.

Often these techniques need to be done in a hospital by specialized personnel. Major limitations include the problem of scheduling frequent analysis with high potential for human error when the machine is not entirely automated.

For the above reasons, the medical community has been looking for fast and reliable screening techniques that can work in an unconstrained environment, with less intervention of specialized, costly physicians. One approach is to identify easy-to-compute indicators that reflect essential health conditions. Common indicators have been used by physicians and nutritionists to specify the human body mass and fat ratio. Less conventional measures have been used in other areas like pharmacology, for drug rate estimation [83], and recently for mortality prediction [234].

1.3.1 Body Mass Index

Body mass index (**BMI**) is the primary measure of obesity [198]. It's defined as the ratio between body mass and the squared height:

$$BMI = \frac{\text{Weight}_{kg}}{\text{Stature}_{m}^{2}} \tag{1.1}$$

BMI represents a measure of the body mass with respect to the height, thus serves as an indicator of relative obesity. BMI was explicitly cited by Keys [146] as appropriate for population studies and inappropriate for evaluating an individual [29].

The BMI is not a perfect measure because it does not directly assess body fat. Muscle and bone are denser than fat, so an athlete or muscular person may have a high BMI, yet not have too much fat. But most people are not athletes, and for most people, BMI constitutes a good gauge of their level of body fat [43]. Research has shown that BMI is correlated with the gold-standard methods for measuring body fat [99]. And it is an easy way for clinicians to screen who

might be at greater risk of weight-related health problems [94], [95]. The interest in an index that measures body fat came with increasing obesity in prosperous Western societies. Some researchers now argue that this flawed and overly reductive measure is skewing the results of research in public health. For years, critics of the body mass index have griped that it fails to distinguish between lean and fatty mass (muscular people are often misclassified as overweight or obese). The measure ignores the distribution of body fat, a critical consideration when it comes to health risks.

1.3.2 Body Surface Area

The whole body surface area (WBSA) is the 2D measured surface area of a human body. Accurate determination of the whole body surface area (WBSA) is one topic that has been actively studied over the last century. From the initial estimate of Du Bois and Du Bois in 1916 [83] to recent work [137], and despite many critiques [259], the WBSA has attracted a lot of attention, driven primarily by the variety of its applications. For many clinical purposes, WBSA is a better indicator of metabolic mass than body weight, since it is less affected by abnormal adipose mass [214]. WBSA is used primarily in pharmacology to estimate drug dosage rates [83] since it is proportional to the absorbing rate [141]. WBSA has been used in medicine to help determine dosing rates and strategies for anticancer drugs and radiation dose estimation [89], [270]. The renal clearance is usually divided by the BSA to gain an appreciation of the correct required glomerular filtration rate (GFR) [13], [214]. WBSA is also used to quantify skin burn areas [122]. An assessment of the burned body surface area is indispensable for evaluating whether the patient requires hospitalization for intravenous fluid resuscitation [93]. In [137], [13] the WBSA was used to account for different body sizes in patients with aortic stenosis. Aortic valve area (AVA) is divided by body surface area (BSA) to calculate the indexed AVA (AVAindex). Calculating the surface area is particularly important in plastic surgery [157]

to determine the area of the skin needed. WBSA is used in the fashion industry for size dresses and accessories [162], and in ergonomic design [254].

We believe that **the WBSA coupled with computer vision techniques can successfully overcome the usual problems with BMI, namely the inability to capture the distribution of body mass, and failure to distinguish between lean and fatty mass**. Measuring the WBSA is, however, a problematic task involving the surface estimation of a non-rigid 3D object. The WBSA as the measure of the surface area, differently from the BMI, is a physical attribute. This fact is fundamental, since the WBSA can be measured directly with computer vision techniques, instead of estimating using the pair of weight and stature.

Historically, the only easy way to get this measure (WBSA) is through some empirical formulae that consider just two human body parameters (body weight and stature). The large variety of body shapes, body composition, and race make the use of a fixed formula highly questionable. Thus there has been a continuous stream of efforts to accommodate different individuals [200],[101],[76], [142], [305], [304], [187]. Other recent approaches is to use direct measurements using a three dimensional (3D) whole body scanner [304], [292], [295]. The problem is that such scanners are expensive and costly to run, thus limiting their availability to users. Yu et al. [304] provide more detailed analysis on some of these problems.

Automated measurement of the WBSA that is accurate, cheap, reliable, and convenient to the subject still remains a fundamental challenge.

1.3.3 Body Fat Percentage (BFP)

The body fat percentage (BFP) is a measure of fitness level and is one of the few measurements that can measure a person's relative body composition without regard to height or weight. The BFP of a human or animal is the total mass of fat (Weight$_{FAT}$) divided by total body mass

(Weight$_T$); body fat includes essential body fat and storage body fat.

$$BFP(\%) = \frac{\text{Weight}_{FAT}}{\text{Weight}_T} \tag{1.2}$$

Essential body fat is necessary to maintain life and reproductive functions. The percentage of essential body fat for women is greater than that for men due to the demands of childbearing and other hormonal functions. The percentage of essential fat is $3 - 5\%$ in men, and $8 - 12\%$ in women [134]. Storage body fat consists of fat accumulation in adipose tissue, part of which protects internal organs in the chest and abdomen. The minimum recommended total body fat percentage exceeds the essential fat percentage value reported above [134].

A number of methods are available for determining body fat percentage, such as measurement with calipers, underwater weighing, Whole-body air displacement plethysmography, also called BodPod, near-infrared interactance, or through the use of bioelectrical impedance analysis. There are also some anthropometric methods for estimating body fat, often using a formula relating the body measurements to density [84]. These methods are therefore inferior to a direct measurement of body density and the application of just one formula to estimate body fat percentage. One way to regard these methods is that they trade accuracy for convenience since it is much more convenient to take a few body measurements than to submerge individuals in water.

The chief problem with all statistically derived formulae is that to be widely applicable, they must be based on a broad sample of individuals. The ideal statistical estimation method for an individual is based on a sample of similar individuals. For instance, skinfold estimation methods are based on a skinfold test, also known as a pinch test, whereby a pinch of skin is precisely measured by calipers at several standardized points on the body to determine the subcutaneous fat layer thickness [256]. A skinfold based body density formula developed from a sample of male collegiate rowers is likely to be much more accurate for estimating the body density of a

male collegiate rower than a method developed using a sample of the general population. Since the sample is narrowed down by age, sex, physical fitness level, type of sport, and lifestyle factors. On the other hand, such a formula is unsuitable for general use.

1.4 Human Body Shape Analysis, A New Approach

In this chapter, we have introduced some problems related to human metrology, soft biometrics, and medical science. We have found some common grounds in these disciplines to conclude that a unique approach can be taken. We also realized that there is space for a new method in the analysis and representation of the human body. Fundamental to our approach is the use of computer vision and machine learning techniques, and recent innovations in acquisition devices. In particular, the current data-driven approaches show superior performances and more robust results compared to model-based approach.

One key problem in body shape analysis is the lack of suitable datasets for scalable analysis. An appropriate dataset is expected to meet some key criteria:

- Comprehend accurate measurements as well as related 2D/3D data.
- A considerably large number of individuals.
- Significant diversity in the samples forming the dataset.
- Should be freely available

Despite the actual trend in computer vision, where the amount of data has seen exponential growth (e.g., ImageNet challenge [250]), other fields have only experienced moderate or limited growth. The causes are mainly due to the nature of the data: 3D data is still expensive to acquire and store, and the labeling process is slow and costly. Today datasets meeting these criteria are

still scarce. The CAESAR dataset [244] partially meet these criteria. CAESAR 3D, composed of measurements and 3D mesh of individuals has been used for this kind of analysis but is not free, costing $10,000 and contains only 2300 individuals.

To overcome the limited amount of data, we decided to use synthetic data in the form of skinned mesh models. In particular, we created a virtual environment able to generate virtual subjects, with the critical capability to control the generation process. This special characteristic permits to create a new dataset composed of "virtual" subjects, with anthropometric measurements that resemble a population of real humans. Nonetheless, the virtual environment allows controlling other essential aspects of the new samples, like the automatic labeling of the subjects. In Chapter 2 we introduce the new dataset composed of virtual subjects, and the new flexible and controllable environment for automatic labeling. In Chapter 5 we expand the method to bodies in multiple poses.

We prove the usefulness of the new dataset, designing a computer vision system able to estimate the WBSA from a single viewpoint. The particular approach developed in Chapter 3 is unique and quite intuitive but masks a well-known problem in computer vision: the evaluation of the surface area of a non-rigid 3D object (body) from a single view. According to Marr's information processing [193], the viewer-centered description is also called 2.5D view. This representation is a mid-level representation between the raw *primal sketch*, which is mainly concerned with the description of the intensity changes in the image and their local geometry, and the *3D model*, which is an object-centered representation of three-dimensional objects.

In Chapter 4 we introduce the use of Spectral Geometry (SG), a sub-field of geometry processing, in the computation of a pose-invariant human body description. A common problem in anthropometric measurements is that of pose variation. The human body can assume a large variety of poses, and since it is a non-rigid object, the shape changes significantly between the poses. Often, body measurements involve the computation of circumferences, and rectilinear

measures on ill-defined points (e.g., torso size, breast, etc.). These situations, usually easy for a human, are still quite challenging for a machine. We discuss important relations between SG descriptors, and human body indicators: Body Fat Percentage (BFP), as well as BMI and WBSA. Specifically, we assess the invariance of the spectral descriptors with vertical (constant BMI), and horizontal (increasing BMI) variations of the body mass.

In Chapter 5 we describe an accurate statistical analysis that proves the variability of the anthropometric measurements under pose variations. Using classification and retrieval tasks, we show that the performance of anthropometric measurements degrades with increasing body pose variation. The Chapter also compares the introduced SG techniques with the traditional anthropometric measurements in a typical soft biometric scenario. In this work, we describe a new machine learning architecture able to regress traditional human understandable descriptors (measurements) from abstract spectral features. The approach we take is entirely new for this field, mainly because we want to overcome the limitations and constraints as discussed earlier in this chapter.

In human body modeling, it is of particular interest to find a parametric representation of body variations. For instance, BMI and WBSA have been used to track the body changes of subjects, but suffer from various drawbacks as cited earlier. Considering the Waist-to-Height ratio (WHR), is interesting to analyze the semantic characterization of different subjects. This method can be beneficial to study the space of the body variations: high dimensional, non-linear, difficult to analyze. In Chapter 6 we propose a new generative approach to analyzing geometric body variations. Taking advantage of the recent development in **deep learning** [126] and **adversarial learning** [104], we design a generative model, able to create new bodies comprises in the similar distribution of body measures. The method explores a lower dimensional space learned by an unsupervised adversarial technique. We discover unusual patterns in body variations when adopting different sampling strategies.

In the appendix, we describe some interesting techniques complementing the work reported in the core chapters. Inspired by the WBSA analysis in Chapter 3, and by recent work [274], [275] we build a renderer to create real "views" of the 3D human body(Appendix A). This system is capable of generating millions of views from the subjects in the dataset, permitting us to simulate the human body in a real environment. In Appendix B we report some important proof regarding Spectral Geometry. Although we limited the extensive mathematical framework, we tried to build a substantial background able to explain the 3D body shape.

Chapter 2

VirtualBody: A Virtual Dataset for Body Shape Analysis

2.1 Introduction

The commonly used body shape datasets in soft biometrics and health assessment include: CAESAR [244], MoCap [67] and NHANES [56] datasets. These datasets are interesting resources to study the human body shape in a large variety of applications: healthcare, medical sciences, ergonomic studies, soft biometrics, forensics, etc. A feature common to all three datasets is the possibility to obtain anthropometric measures of the body. However, given the different nature and specific goals of these collections, each focus on a particular problem. CAESAR [244] and NHANES [56] datasets lean more toward clinical applications, with supplemental geometric information (CAESAR 3D). MoCap dataset [67], instead, did not have the measure of the body shape, as it's primary goal. It's focus was to track and detect the motion and pose of the body. The data acquisition is done using an automatic system (Vicon) composed of a set of synchronized cameras that detect the position of reflective markers on the body.

Given the location and numbers of markers on the body, it is possible to extract antropometric measurements from the data. This methodology is definitely useful for body shape analysis, but unfortunately the available datasets do not represent a population of subjects with sufficient statistics. This is a common problem, since collecting a significant amount of data from different subject classes: gender, race, ages, health conditions is expensive, time-consuming, and sometimes challenging.

The acquisition of 2D/3D data from subjects is often a challenge. Fast and reliable methods for 3D acquisition only became available recently. A common issue in data acquisition from humans is the subject privacy. Unfortunately, data-driven machine learning algorithms need training data from a large number of body shapes with significant diversity.

Since collecting this large amount of data is expensive, and infeasible for a research laboratory, we decided on a **virtual** approach. We propose a generative-based framework where elements are **virtual subjects**, associated with computer vision and computer graphics techniques for the human body analysis. Under this framework, we generate a large number of virtual subjects (3D mesh data) that can capture variations in body shape and body size due to gender, race, and age. This virtual population needs to capture the statistical attributes of a real population with all the possible body shapes. The generation of synthetic data is not new in computer vision (e.g. [266], [45]). However, our dataset is unique in its focus on human body shape, its size, and diversity.

2.2 Shape Semantics

In this chapter, we focus on the generation of meaningful data where the **semantic** information is the most valuable asset. **Semantics** (e.g., meaning or functionality in a given context) is still an overlooked feature for data in shape analysis. This is partly due to the lack of methods

for automatic extraction of semantic content from digital shapes, otherwise known as semantic annotation, and partly to the evolution of research on **shape modeling**, which in the past years was highly focused on the geometric aspects of shapes. The principal benefit of generating a virtual dataset is the capability to associate powerful **semantic features** to the generated data. This peculiarity will allow us in the next chapters to develop powerful techniques able to "learn" features and concepts not usually available or not easy to annotate in real data. When we talk about **semantics** related to the human body we can define different levels of features, for various applications.

Due to the nature of our study we are interested in all the quantities, and subject conditions that influence the visual appearance. Thus, the number of features can be quite large, since the human body can assume different shapes with age, gender, race, health status, and body pose. It is a non-rigid object. Although there has been some significant work to lower the complexity in the analysis of body shape, the representation of the human body is still an open problem. The approach that seems most promising is to use a parametric model that can describe the body using a pool of parameters. These parameters can be considered as many semantic features (anthropometric measurements), or as unique features when a set of parameters define a subject as fat, or lean. In our approach, we use a parametric model to define each subject. Then a graphics engine will create the final mesh given the model parameters and body pose data. Here, we focus more on shape analysis without motion, which is the most common situation in a physician's clinic. However, the same framework can be used to track and analyze subjects in different poses.

2.3 Related Work

Related work can be found in body modeling, human body measurements, and computer graphics. However, although there are many published works in these areas, just a few present a dataset with a significant number of subjects. Traditionally, there have been many models to represent the human body. From 1D structure living in 3D space, skeleton-based, to 2D and 3D models. To restrict our attention to the most related and recent work, we will focus on 3D non-rigid body models.

2.3.1 Datasets

SCAPE [7]: Scape is one of the most popular non-rigid parametric models is the SCAPE method, but the meshes used to train and test the model has been released as 3D dataset. The dataset contains 71 registered meshes of a particular person in different poses. With the original dataset has been published the morphs of the template models, the scape-completion of each scan, and the correspondences between the template and each scan. Although the mesh description is very accurate, the number of subjects in the dataset is not statistically significant for the human body shape analysis.

CAESAR 3D [244]: The Civilian American and European Surface Anthropometry Resource Project is an extensive 3D database including measurements from the entire North American population sample (2400 male and female subjects, aged 18-65) including demographics. This database is the first to include 3D model scans together with traditional 1-D measurements. Scanned poses are: standing, relaxed seated, and coverage poses. In addition, the database contains 40 traditional (1-D) anthropometric measurements done with a tape measure and caliper. This dataset is the most complete 3D dataset from real scans available. However, this dataset is

not free, and for human shape analysis can lack significant variability in the shape population.

TOSCA [36]: Bronstein et al. created a dataset for 3D shape retrieval. Shape retrieval focuses on the design of a shape descriptor or signature, which captures the unique properties of the shape, and is invariant to a certain class of transformations. In rigid shape analysis, common transformations are rotation and translation. In shape retrieval problems, the number of transformations is more vast: scale, missing parts, different sampling and triangulation. The database contains a total of 80 objects, including 11 cats, 9 dogs, 3 wolves, 8 horses, 6 centaurs, 4 gorillas, 12 female figures, and two different male figures, containing 7 and 20 poses respectively. Since is not composed only of human figures, it is very limited for human body shape analysis.

SHREC'10 [37]: Bronstein et al. extended the Tosca dataset adding more challenges: robust large-scale retrieval, correspondence, and features detection and description. The database contains a total of 148 objects, including 9 cats, 11 dogs, 3 wolves, 17 horses, 15 lions, 21 gorillas, 1 shark, 24 female figures, and two different male figures, containing 15 and 20 poses respectively. Unfortunately, the number of human subjects had only a slight increase.

FAUST[31]: Bogo et al. takes advantage of the new scanning technologies to create a new dataset of human bodies. This work is manly focus on surface registration. The registration is particularly challenging for non-rigid and articulated objects like human bodies. The authors address the registration problem with a novel mesh registration technique that combines 3D shape and appearance information to produce high-quality alignments. The new FAUST dataset contains 300 scans of 10 people in a wide range of poses together with an evaluation methodology. This dataset present data of real subjects in a variety of poses, but it is still very limited in the number and diversity of subjects.

NHANES [56]: The National Health and Nutrition Examination Survey (NHANES) is a program of studies designed to assess the health and nutritional status of adults and children in the United States. Findings from this survey will be used to determine the prevalence of major diseases and risk factors for diseases. This dataset is composed only of numerical values in the form of tables. There are no images or 3D data of the subjects, although, it presents an enormous source of information on health, habits, and morphology of the subjects as anthropometric measurements. This dataset, although it appears not very relevant from a computer vision viewpoint, it will be very useful for the statistical analysis of the population, and as we will see later, for generating virtual subjects.

2.3.2 Models

Below we describe some 3D skinned body mesh models:

SCAPE [7]: One of the most popular non-rigid parametric models is the SCAPE method [7]. The SCAPE method is a data-driven method for building a human body model that spans variations in both shape and pose. The method is based on a representation that incorporates both articulated and non-rigid deformations. Learning the model is constituted by two operations: learning a pose deformation model from a subject with multiple poses, and learning a shape model from many subjects with a neutral pose. The decoupling of shape and pose deformations in the SCAPE model has a significant limitation: 3D meshes of different individuals can change similarly for the same pose change. Various efforts have been made to improve accuracy and constraints of the SCAPE model. Chen et al. [61], proposed a new improved body model promising more accuracy and faster fitting time from real data, by exploring a tensor decomposition technique. This decomposition permits to model the deformation as a joint function over both shape and pose parameters to preserve the dependency between them.

BlendSCAPE [127]: Hirshberg et al. confront the ill-posed problem of joint modeling and registration together. The solution is minimizing a single objective function, obtaining high-quality registration of noisy, incomplete, laser scans, while simultaneously learning a highly realistic articulated body model. This model drammatically improves robustness to noise and missing data. Since the model explains a corpus of body scans, it captures how body shape varies across people and poses.

Delta [30]: Delta is a method to estimate the 3D geometry and appearance of the human body from a monocular RGB-D sequence of a user moving freely in front of the sensor. RGB-D data in each frame is aligned with a multi-resolution 3D body model in a coarse-to-fine process. Then using multi-frame geometry and image texture, obtain accurate shape, pose, and appearance information could be extracted despite unconstrained motion, partial views, varying resolution, occlusion, and soft tissue deformation. The novel body model has variable shape detail, allowing it to capture faces with a high-resolution, deformable head model and body shape with lower-resolution.

Dyna [227]: Dyna focused on soft tissue deformations, like those of real people, using a high-resolution 4D capture system. The method accurately registers a template mesh to sequences of 3D scans. Using a powerful acquisition system, it's possible to acquire over 40,000 scans of ten subjects. At this frame rate, the system can learn how soft tissue motion causes mesh triangles to deform relative to a base 3D body model. The Dyna model uses a low-dimensional linear subspace to approximate soft-tissue deformation and relates the subspace coefficients to the changing pose of the body. Dyna models how deformations vary with a persons body mass index (BMI), producing different deformations for people with different shapes. Dyna realistically represents the dynamics of soft tissue for previously unseen subjects and motions. Besides the good results, the proposed work is still based on a small number of subjects, already

released in the FAUST dataset [31]. As explained earlier, there is still some confusion on the use of BMI as a relative measure of body fat.

SPML [313]: Zuffi et al. propose a new 3D model of the human body that is both realistic and part-based. The model represents the body by a graphical model in which nodes of the graph correspond to body parts that can independently translate and rotate in 3D as well as deform to capture pose-dependent shape variations. This model defines a "stitching cost" for pulling the limbs apart, giving rise to the stitched puppet model (SPM).

Hasler et al. [116]: focus their attention on the generation and animation of realistic humans. In this work Hasler proposed a unified model that describes both, human pose and body shape, permitting to accurately model muscle deformations not only as a function of pose but also dependent on the muscle bulging of the subject. The proposed model is based on statistical analysis of over 550 full body 3D scans taken of 114 subjects. All subjects are measured with a commercially available impedance spectroscopy body fat scale and a medical grade pulse oximeter. Although the dataset includes information on anthropometric measurements and body composition, 114 subjects is still a small number to learn a complete population. However, this data will be useful in the future when we want to use real data.

2.3.3 Methods

Lie Bodies [96]: Freifeld et al. show how to characterize the set of all possible deformations in a human body. This unique approach, grounded in differential geometry, provides an elegant approach to the representation of the spaces of subjects from different classes (gender, weight, etc.). In this case, each subject lies on the surface of a lie manifold embedded in R^3. The deformations applied to each body mesh form a Lie group, and the authors proved that all the

rules are valid for this environment. Freifeld et al. extended this framework in [97] combining transfer learning and parallel transport to improve the learning of datasets with missing subjects. Black's framework is based on triangular meshes of subjects contained in the CAESAR 3D dataset [244]. This dataset includes mainly Caucasian subjects (Europe and North America) with minorities as Asian and Afro-American.

Kinect @Home [295]: Weiss et al. proposed one of the first methods to acquire 3D structure of a body in a more relaxed environment. This method, taking advantage of the new Microsoft Kinect [285] obtained good results. However, the joint optimization involved in the registration and fitting of the 4 point clouds on the body model makes the system extremely slow (40 min for one subject).

OpenDR [183]: Loper et al. improve the acquisition system with a new technique. The inverse rendered technique attempts to take sensor data and infer 3D geometry, illumination, materials, and motions such that a graphics renderer could realistically reproduce the observed scene. Renderers, however, are designed to solve the forward process of image synthesis. To invert the process, the authors propose an approximate differentiable renderer (DR) that explicitly models the relationship between changes in model parameters and image observations.

MoSh [184]: is a marker-based motion capture (MoCap) system. In the last decade, these systems have been widely criticized as producing lifeless animations. The authors argue that important information about body surface motion is present in standard marker sets but is lost in extracting a skeleton. This approach automatically extracts this detail from mocap data, estimating body shape and pose together using sparse marker data by exploiting a parametric model of the human body. In contrast to previous work, MoSh solves for the marker locations relative to the body and estimates accurate body shape directly from the markers without the

use of 3D scans; this effectively turns a mocap system into an approximate body scanner. MoSh is able to capture soft tissue motions directly from markers by allowing body shape to vary over time.

2.4 3D Body Model and Virtual Body Framework

In our work, we take advantage of a different body model. Makehuman [16] (MH) is an open-source 3D computer graphics application, designed for the prototyping of photorealistic humanoids to be used in 3D computer graphics. MH takes advantage of 3D morphing technology. Starting from a (unique) average human base mesh, it can be transformed into a great variety of characters (male, female, African, Caucasian, Asian, adult, kid, etc.), using a linear interpolation of different target models. Using this technique, one can reproduce different characters with very different body shapes. The model has two types of parameters:

- **macro parameters**: stature, weight, gender, ethnicity and muscularity (fat / muscle ratio).
- **micro parameters**: body part measurements (waist circ., torso, thigh circ., etc.).

Macro and micro parameters constitute the parameter sets that define each subject. MH is specifically designed for modeling virtual humans as characters in virtual reality and gaming, with a simple and complete pose system that includes the simulation of muscular movement. The parameterized model and the extreme simplicity in creating characters make MH a handy tool for our environment.

However, our goal is to generate an entire population of thousands or more of individuals with some specified statistical distribution. To realize this task, MH was not directly usable, since it was built to design game characters one at the time. To overcome this limitation we

develop a new plugin able to take advantage of MH graphics engine.

2.4.1 Generation of Virtual (Synthetic) Humans

MakeHuman has been used before to create a dataset of realistic human bodies. The main applications have been in the generation of a human population for bed fitting [286], for learning a random forest in a computer vision system [45],[44], and on camera positioning [224]. All these works, although, do not present an efficient technique to generate a population of subjects with parameter variance similar to a real population.

The MH parametrized model can be stored efficiently in a file containing the parameter values. Useful available parameters include skin texture and clothes as part of the model. In our work we included the **Caucasian** skin texture in all the subjects in the datasets as shown in Figures 2.6 and 2.7, but we have available African, and Asian skin textures too.

We developed a plugin able to read a set of parameters for each subject from a file, automatically create the desired mesh structure and save it in the right format. The pipeline of the generation process is shown in Figure 2.1. With this plugin, we can create thousands of bodies in a relatively small time ($\sim 2h30min$ on a quad core CPU for 20000 subjects). Since a mesh is a real 3D object with physical measures, we implement a semantic features generator. We take advantage of the MH measuring tool library to measure the generated mesh and store them in a table of body measurements in NHANES [56] style. This tool is the critical part of the plugin because it allows us to automatically store all the parameters, measurements, and semantic information. The plugin has been designed while considering the different possible scenarios where the generated data can be used. We describe two possible situations (or datasets), but the plugin can be re-configured easily.

- **Completely random virtual dataset** (20000 subjects), called **Virtual Random dataset**.

- **NHANES-based dataset** [56] called **virtual NHANES dataset** (12500 subjects).

Table 2.1 provides some statistical data on the two introduced datasets.

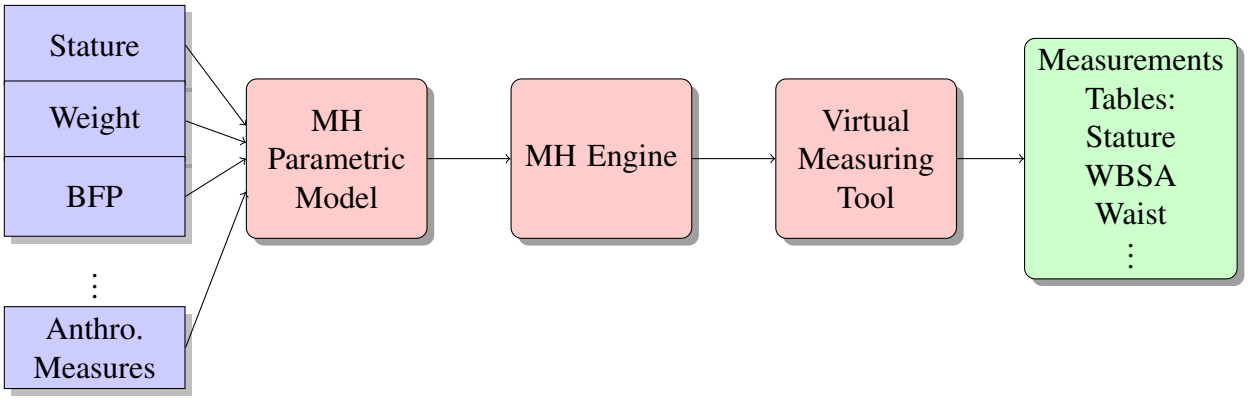


Figure 2.1: VirtualBody Method Pipeline.

Virtual NHANES dataset

This Virtual dataset has the goal to mimic a real human population for health assessment studies. Since we can easily and freely obtain datasets with body part measurements (CAESAR 1-D [244], NHANES [56]), we decided to use these measurements to build the respective virtual subjects mesh. We use the subject measurements available from the National Health and Nutrition Examination Survey III (NHANES III) dataset (ages 10-85) [56]. Using the body measurements from NHANES we generated the corresponding set of macro and micro parameters for our model, and subsequently the triangular meshes and the annotation table. The process runs automatically, reading the subject measures from the NHANES tables and generating the outputs without human intervention. The parameters used for the generation are: gender, age, height, race, breast size, upper leg height, upper arm length, upper arm circumference, thigh circumference, and waist circumference. MH represents all the macro parameters and some micro parameters as a normalized value between 0 and 1. For some of these parameters, we know the range used by MH, in which case we can recover the real measure. For some, we do not.

We decide to allow these parameters to be variable in the data range. The first reason for doing so is that since MH use a normalized weight, it could be misleading to normalize the NHANES weight with the MH range. The second reason lies in the targeted experiments for this dataset: study the health indicators related to body shape. Changing weight and muscle ratio, but keeping the other parameters fixed is like varying the body mass of the subject. But at the same time, by varying the muscle/fat ratio, we obtain a *fat version* and a *skinny version* of the same individual. This is very interesting since it can be used to learn how the WBSA change with the respective variations in weight and muscle/fat ratio. In fact, analyzing NHANES dataset [56], we discover that many individuals are very similar, and we couldn't get a larger and continuous shape variation. Thus, we generated a population composed of a total of 12500 subjects for the Virtual NHANES dataset: 25 meshes for each subject, for 500 original subjects. As a side note, we specify that the physical measurements obtained as output measuring the mesh are real values in cm, and these are part of the values stored in the output table.

Virtual Random dataset

The Virtual NHANES dataset is aimed at mimicking a real population with some interesting augmentation. However, actual data often come in the form of subjects with random statistics that can assume some distribution. This kind of data can be very challenging because there can be somebodies that could be hard to find in a real population. We designed a new modality in the plugin able to generate a random population with a specified distribution. The result is our **Virtual Random** dataset. To create this dataset, we allow the plugin to generate random values for the following macro parameters (stature, gender, race, weight, and muscle ratio). To increase the variability of the obtained bodies, we generated these parameters using a uniform distribution rather than a normal distribution. Real population distributions for the parameters is close to a normal distribution, however using a uniform distribution guarantees a higher number

of subjects at the extremes of the possible ranges. In fact, as shown in [4], the WBSA of subjects at the extremes (e.g., kids, and very obese subjects) can create significant problems in the body shape analysis. However, to avoid the creation of subjects that are too dissimilar from real human bodies we restrict the randomly generated parameters to more realistic intervals. The Virtual Random dataset is important in evaluating the performance of body shapes analysis methods at extreme body shapes and body sizes.

2.5 Results: Virtual dataset

Figure 2.2 shows the mesh model for one of the generated subjects, while Figure 2.3 shows the distribution of the WBSA in the datasets. Figures 2.4, and 2.5 show the distributions of the stature and Waist-to-Stature-Ratio (WSR) for the Virtual NHANES dataset, and Virtual Random datasets. Virtual NHANES values are in the range of a real population since the measurements are extracted from the NHANES dataset. Virtual Random, as defined above, has a higher number of unusual subjects. In fact, from the histogram on the left of Figure 2.4 the tails of the histogram are longer than the Virtual NHANES. However, although highly unlikely, is not impossible since there is a record of a man that is 2.73 meters tall [82]. Samples of males and females for different muscle/fat ratios in the dataset are shown in Figures 2.6 and 2.7. MH defines a texture for a given subject based on gender, age, and races. It is also possible to add some other structures such as short or long hair. However, this feature has not been used. Table 2.1 shows the compositions of the generated datasets. We have included information on the EORTC (European Organization for Research and Treatment of Cancer [270],[253]) dataset for comparison. For the generated datasets the WBSA is computed from the original mesh. For the EORTC, the WBSA is computed using the traditional formulae. The Virtual Random dataset has a notably larger variance containing many varieties of subjects. The Virtual Random dataset

includes subjects that are hard to find in the modern population (notice that in Figure 2.3, left, there are subjects with WBSA approaching $400dm^2$!). The EORTC has an average WBSA higher than the Virtual NHANES. Since the EORTC considers cancer patients, it is composed almost exclusively of adults. Our dataset instead is comprised of a large variety of ages.

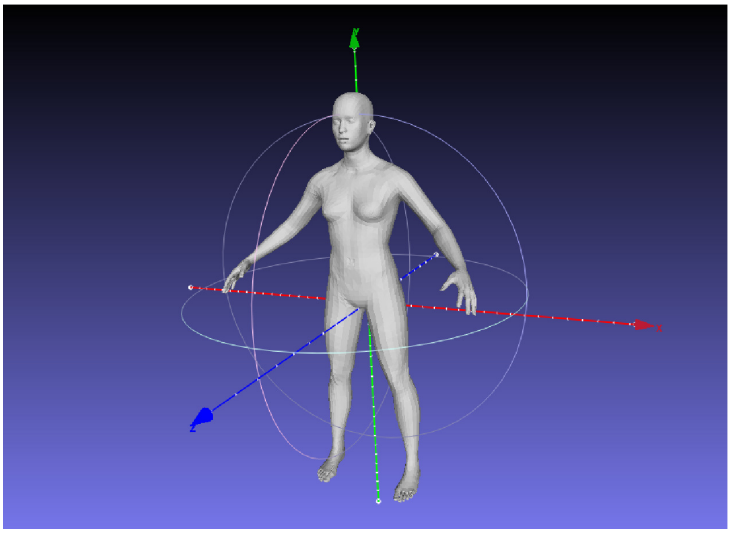


Figure 2.2: MakeHuman mesh model.

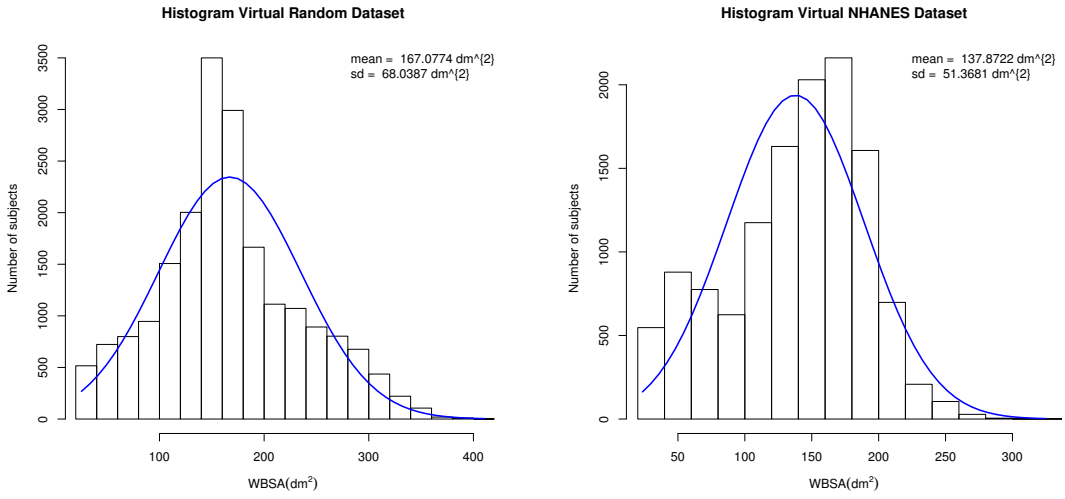


Figure 2.3: Distribution of WBSA in the proposed datasets:Virtual Random (left) and Virtual NHANES (right).

Table 2.1: Statistics on the datasets.

	Virtual NHANES	Virtual Random	EORTC
Total subjects	12500	19995	3000
Males	6348	10049	
Females	6152	9946	
Kids (≤ 15) yrs	4123	5786	
Adults (> 15 yrs)	8377	14209	
Small ($H \leq 130$ cm)	3172	14209	
Normal ($H = 130 - 200$ cm)	9213	12449	
Big ($H > 200$ cm)	76	4612	
Ages	$10 - 85$	$12 - 70$	adult
Mean WBSA (dm^2)	137	167	173
SD WBSA (dm^2)	51	66	

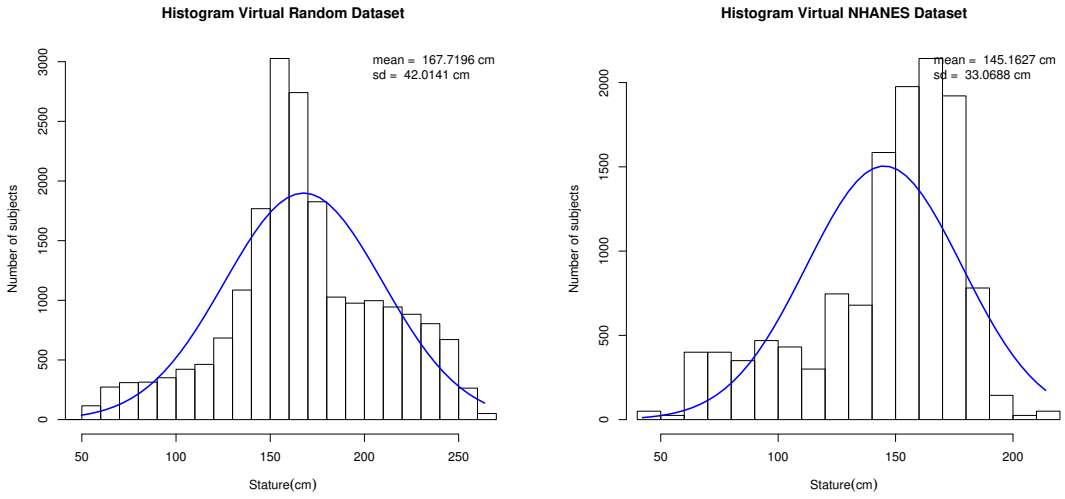


Figure 2.4: Distribution of the Stature in the proposed datasets:Virtual Random (left) and Virtual NHANES (right).

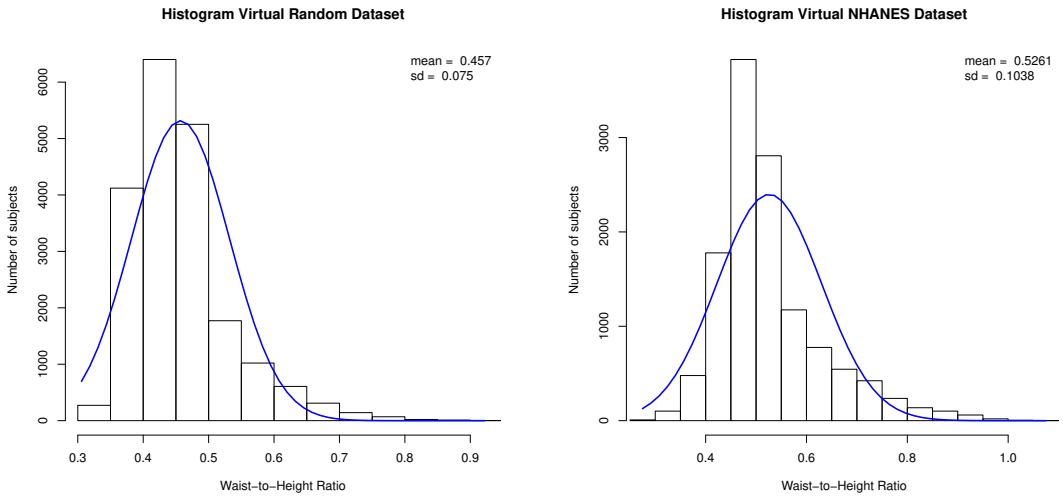


Figure 2.5: Distribution of the WSR in the proposed datasets:Virtual Random (left) and Virtual NHANES (right).

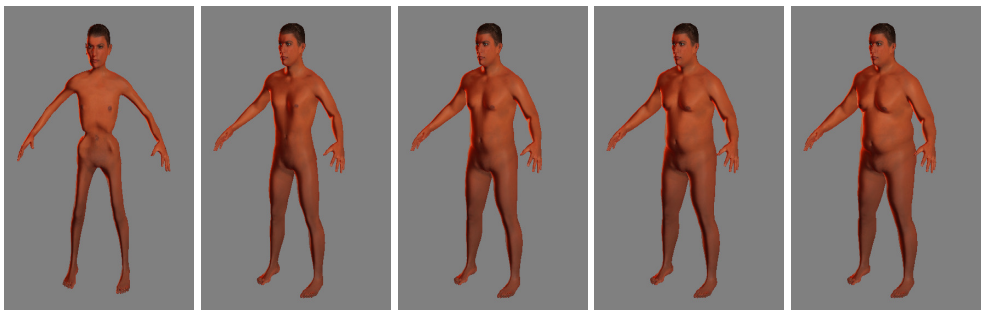


Figure 2.6: Male subjects in Virtual NHANES dataset.

Figure 2.7: Female subjects in Virtual NHANES dataset.

Chapter 3

Whole Body Surface Area Estimation

3.1 Introduction

Accurate determination of the whole body surface area (WBSA) is one topic that has been actively studied over the last century. In section 1.3.2 we introduce the importance of the accurate determination of this critical indicator. In this chapter, we propose a virtual framework able to study the whole body surface area **WBSA**. Fundamental of this chapter is the belief that since the WBSA is a geometric measure, it can be estimated more accurately with computer vision techniques, rather than with weight and stature. The WBSA computed with the usual formulae [83] suffer from the same problem as BMI: it doesn't consider the body composition, but the error from the real value has a different effect. However, the WBSA computed with computer vision techniques can easily overcome the usual problems with BMI, namely the inability to capture the distribution of body mass and inability to distinguish between lean and fatty mass. Because fat and lean mass density are way different, then higher Body Fat Percentage (BFP) will have a different effect on the visual appearance.

Historically, the only easy way to get this measure (WBSA) through some empirical formulae

that consider just two human body parameters (body weight and stature). The large variety of body shapes, body compositions, and races makes the use of a fixed formula highly questionable. Thus there has been a continuous stream of efforts to accommodate different individuals. Another recent approach is to use direct measurements using a three dimensional (3D) whole body scanner. The problem is that such scanners are typically costly, costing hundreds of thousands of dollars, and have to be used by trained personnel, thus limiting their availability to users.

3.1.1 WBSA: Measurements and Estimation

The conventional methods for WBSA calculation are through some well-known formulae. The most widely used formula for WBSA calculation is the one devised by Du Bois and Du Bois in 1916 [83]. Molds of plaster of Paris for nine subjects were cut into small pieces in an attempt to measure the two-dimensional surface area of the skin. Each subject's body/skin surface area was then calculated, and Du Bois and Du Bois determined that WBSA was related to stature and weight by the formula: $0.007184 \times W^{0.425} \times H^{0.725}$ [83], where W is the weight (in kg), and H is the stature (in cm) of the subject. Notably, this formula was derived from 9 subjects only, one of whom was a child. Since the bodies of the subjects studied in the middle of the First World War are unlikely to be similar to the patients of the modern society, Mosteller proposed a new calculation of WBSA in 1987 [200]. This formula is a modification of the WBSA equation by Gehan and George [101].

Today there are many studies related to the verification of meaningful differences between WBSA measurements taken using a whole body three-dimensional (3D) scanner (criterion measure) and the estimates derived from each WBSA equation identified from systematic reviews [76], [142], [305], [304], [187]. In these studies, the 3D scanners used are often cumbersome and slow and have to be operated by specially trained personnel. The formulae are still in use,

but many corrective factors are appearing to adapt the formulae to today's special cases (e.g., very obese people) [292],[255],[182], or race [5], [305]. Verbraecken et al [292] examined the WBSA based on Mosteller's formula in normal-weight (BMI, $20 - 24.9$ kg/m), overweight (BMI, $25 - 29.9$ kg/m), and obese (BMI, > 30 kg/m) adults (> 18 years old) in comparison with other empirically derived formulae. With obesity, weight increases without a proportional rise in stature. Consequently, it is possible that the WBSA-predicting equations, which include stature coefficients, could systematically miscalculate WBSA for obese patients. Because many clinically essential measurements are indexed to WBSA, systematic errors in WBSA estimation can adversely affect the clinical care of obese patients. Similarly, [255] and [182] showed that the well known WBSA formulae (DuBois and Dubois) fail to accurately predict the WBSA at the extreme of the normal weight range (10-80 kg). Different scenarios are analyzed in [5],[14],[305] each requiring a different modification of the basic WBSA formula.

Measurements using body scanner

An alternative to the use of WBSA formulae is whole-body 3D scanning. There are three significant issues with the 3D laser scanners: cost, speed, and physical space requirement. Classic 3D laser scanners use a laser beam to illuminate the surface. At the same time, a receptor registers the beam distortion on the surface and computes the respective depth. The beam needs to cover all the space of the surface, and it takes time to do so. The process requires that the object is almost immobile and small movements can cause errors in the reconstruction. Modern laser scanners are fast enough to avoid this distortion, but still, require a large room to contain the device.

The result of the scanning operation is usually "raw" data in the form of a 3D (x, y, z) point cloud. To reconstruct the mesh surface from the raw data, a surface reconstruction algorithm has to be applied. Without the face information, it is not possible to relate the vertices to a

face and thus compute the area of the surface. The 3D data, after surface reconstruction, is completed by other information than (x, y, z) points. The triangles tessellation, for instance, fits many little triangles every 3 points. Then the calculation of the whole body surface area is reduced to a simple summation of the areas of all the triangles composing the mesh. This solution, unfortunately, is not as reliable and efficient as it looks. Key challenges in 3D body scanning include occluded areas [304], body parts registration [292], [295], device complexity and portability. Yu et al. [304] provide more detailed analysis on some of these problems.

RGB-D Cameras

State-of-the-art RGB-D cameras are getting smaller, more accurate, and cheaper. This class of devices is led by the well known Microsoft Kinect for XBox [199]. This device permits to acquire 3D data with a simple home setting. However, Microsoft Kinect [199] is far to be portable, since it requires a minimal pc to work, and the power requirement for both is not negligible. To overcome these drawbacks we designed as a lateral project a structured light system composed of a smartphone and a low power pico projector [221]. We tested the system on the face acquisition task. Although the great performance we were not able to use this device for the impossibility to recruit a sufficient number of subjects.

A surprising result was reported by Weiss et al. in [295]. With only one device in a home setting, they develop a system capable of reconstructing the 3D mesh using four views of the subject. This methodology avoided the use of cumbersome 3D scanners but has some limitation. Acquiring many different views of the subject requires a robust registration process. Moreover, the registered views are used to fit a model, in this case, the SCAPE model [7], to build the parametrized body model. This process, unfortunately, is still computationally expensive and requires a lot of time. The method in [295] requires almost 1 hour to reconstruct the body model. Recently Loper et al. [183] introduced an innovative inverse rendering framework able to speed

up the registration process taking advantage of the modern GPU architecture. The method is significantly faster than [295] with almost the same accuracy, but more prone to errors in the differentiation process if the environment is not well constrained. A challenging problem for the 3D scanner methods, and unfortunately for RGB-D devices, is how to measure occluded areas.

3.1.2 The Problem

The two major streams of work on WBSA (corrective factors for the formulae, and 3D acquisition techniques), have a common problem: both require trained personnel. In fact, the standard WBSA and BMI calculations use prediction equations which are accurate only for patients similar in size to the original study subjects. Using formulae can be apparently more natural in the traditional way that physicians evaluate a subject through weight and height. However, this estimation misses a fundamental component: the body composition. Consider the behavior of the Body Mass Index (BMI) between athletic and overweight subjects. Both have a BMI greater than 25, but one is a healthy athletic subject while the other is an obese subject. This index, unfortunately, is not capable of distinguishing subjects with different body fat percentage. This fact is a common problem in measuring the radiation dose estimation [89] for obese people, where a wrong surface area estimate will create an underdosage of the treatment. To avoid this miscalculation, only trained personnel can establish when a predicting equation is sufficiently accurate or when to use a corrective factor for the given subject. At the same time, classical 3D body scanners, which give a better estimate, cannot be used without supervision either. The use of trained personnel, which can be expensive, could lead to human errors, and is not always feasible, like in an auto-assessment scenario. A simpler, faster and more reliable method to determine the WBSA could provide some significant advantages. Moreover, the formulae have some validity issues with young subjects (< 15 years old) [117], the obese [292], and race di-

versities [14], [5]. Finding new variations or corrective terms for the formulae is very expensive because using the old-fashioned technique with wraps and molds of plaster of Paris or using the modern 3D scanners will require the finding of these subjects, and then spending more time on the measurement process. Unfortunately, using standard 3D datasets such as CAESAR 3D [244] does not solve the problem. In fact, these datasets are limited in subject diversity, and using datasets from different countries can be a solution, but at a cost, and they are not always available.

3.1.3 Virtual Environment

Given these multiple problems, we decided to approach the WBSA calculation with an unusual methodology for this area. Our goal can be summarized with the following idea. *Using a simple Kinect device we want to obtain an accurate estimate of the WBSA for any given person regardless of differences in gender, race, obesity, with the subject merely facing the device without the supervision of trained personnel.* We want to use just one device that can acquire only one view of the subject, simplifying the setting required for accurate estimation, and making possible the precise estimate in a home setting. The device will acquire just the visible portion of the body (View Body Surface Area: VBSA), and a subsequent prediction stage will reconstruct the overall WBSA.

To study this problem, traditional computer vision, and medical trial consider the acquisition of a dataset with real subjects. We decided to avoid this costly solution for a more cheaper solution involving a Virtual clinic.

To obtain the same result of a real 3D acquisition process, like in a clinic, we need to simulate the acquisition process. This stage constitutes the main part of the system, able to reconstruct one view of the body (the side that has been viewed by the camera) from the whole mesh immersed in a virtual 3D room. Analyzing 3D data from a single point of view, usually

away from the surface increase the complexity, since only the visible part of the body can be acquired and analyzed. However, using a virtual environment and virtual subjects constitutes a considerable advantage because we can control at the same time the distortion caused by the acquisition process and the high variability of body measurements when acquired by a non-contact device.

With this setup we seek to find the relationship between the surface area computed from one single viewpoint (we call it view body surface area, VBSA) and the whole body surface area (WBSA) as a function of the camera position through different body shapes. Learning this relation will be extremely useful since we can predict the WBSA of the subject from just one shot.

We targeted a classical physician's office setup with the subject in front of the device. Although designed for WBSA estimation, the presented framework can be used to study a more general problem, such as the behavior of the WBSA, or other geometric measures, in a more unconstrained scenario like video surveillance environment.

In this setting, the position of the body with respect to the camera, the body pose, the camera intrinsic parameters and camera lens distortion all play a huge role in the final measure. The proposed Virtual Environment can tackle all these parameters in one unique model capable of computing the WBSA from the single view (one 3D camera cannot measure the WBSA of the body in one single shot).

3.2 Methods

The method is based on a Virtual camera setup with the subject at the coordinate origin and the camera free to move on the surface of a sphere with the same origin. For this project, we're using the virtual dataset developed in Chapter 2. In the subsequent section, we will analyze the

dataset used, the algorithm behind the virtual camera, and the prediction algorithm.

3.2.1 Dataset

We use the Virtual random and Virtual NHANES datasets developed in Chapter 2. These two datasets contain the WBSA information for each subject computed in the mesh generation phase. For this experiment we don't use the RGB texture because the measure is totally geometric.

3.2.2 Virtual Camera

In computer graphics, a virtual camera system aims at controlling a camera or a set of cameras to display a view of a 3D virtual world with the purpose to show the action at the best possible angle; more generally, they are used in 3D virtual worlds when a third person view is required There are mainly three types of camera systems: fixed camera systems, tracking cameras, and interactive camera systems. Our system can be considered an interactive camera system.

There is a large body of research on how to implement a camera system [273]. Our situation, however, is a bit different than usual virtual camera setup. Given a 3D environment, simple rendering techniques can create different views of the object, containing only 2D information, as in a standard RGB picture. We need to recover, instead, the 3D information from the given camera position, the same result obtained from RGB-D devices, like Microsoft Kinect [199]. This kind of projection has to be based on ray casting technique [248].

Ray casting [248] is the most basic and popular of many computer graphics rendering algorithms that use the geometric algorithm of ray tracing [296]. Ray tracing-based rendering algorithms operate in image order to render three-dimensional scenes to two-dimensional images. Geometric rays are traced from the eye of the observer to sample the light (radiance)

traveling toward the observer from the ray direction. The speed and simplicity of ray casting comes from computing the color of the light without recursively tracing additional rays that sample the radiance incident on the point that the ray hit. This eliminates the possibility of accurately rendering reflections, refractions, or the natural falloff of shadows; however all of these elements can be faked to a degree, by creative use of texture maps or other methods.

The ray tracing technique implemented works in the follow modality. We define a frame composed of "pixel" of the same resolution of the camera model. For each "pixel" we shoot a ray in the direction of the camera. The direction of this ray is due to the intrinsic parameters of the camera (see camera model 3.2.2) and the orientation (extrinsic parameters). For each ray, we compute the point in space (x,y,z) intersection between the ray and the object surface. The obtained point of cloud is the visible part of the object. Rays without intersection are set to zero, following the convention of an organized point cloud. In the next sections, we exploit the role of the camera calibration in the ray casting algorithm, the computation of the surface area, and the VBSA ground truth.

Camera model

The ray casting method gives only the framework to reconstruct a view given the position of the camera. To simulate a real camera we need to add to the ray casting algorithm the camera lens characteristics: intrinsic parameters [114]. We use the pinhole camera model to describe the image acquisition process, which is largely employed to parametrize a large number of cameras. The pinhole camera model defines the geometric relationship between a 3D point and its 2D corresponding projection onto the image plane. This geometric mapping from 3D to 2D is often called a perspective projection. We denote the center of the perspective projection (the point in which all the rays intersect) as the optical center or camera center and the line perpendicular to the image plane passing through the optical center as the optical axis. Additionally, the

intersection point of the image plane with the optical axis is called the principal point. The pinhole camera (see Figure 3.1) models a perspective projection of 3D (X, Y, Z) points onto the image plane (x, y), and can be described as follows:

$$(X, Y, Z)^\top \xrightarrow{Projection} (x, y)^\top$$

The equations of perspective projections are given by

$$x = f\frac{X}{Z} \qquad y = f\frac{Y}{Z} \tag{3.1}$$

where f is the focal length of the camera, i.e., the distance between the image plane and the pinhole.

Intrinsic/Extrinsic Parameters. The complete camera model can be represented with the following relation.

$$\lambda \underbrace{\begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix}}_{\boldsymbol{x}'} = \underbrace{\begin{bmatrix} fs_x & fs_\theta & o_x \\ 0 & fs_y & o_y \\ 0 & 0 & 1 \end{bmatrix}}_{K} \underbrace{\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}}_{\Pi_0} \underbrace{\begin{bmatrix} R & \boldsymbol{T} \\ \mathbf{0}^\top & 1 \end{bmatrix}}_{g} \underbrace{\begin{bmatrix} X_0 \\ Y_0 \\ Z_0 \\ 1 \end{bmatrix}}_{\boldsymbol{X}_0} \tag{3.2}$$

The matrix Π_0 is the canonical projection matrix. The matrix K consists of the intrinsic parameters of the camera. Here f is the focal length of the camera, s_x and s_y give the relative aspect of each pixel. o_x and o_y specify the coordinates of the image center. s_θ is the skew in the shape of the pixel, i.e., its deviation from an axis-aligned rectangle. The matrix g defines the pose of the camera. The elements of g constitute the extrinsic parameters of the camera

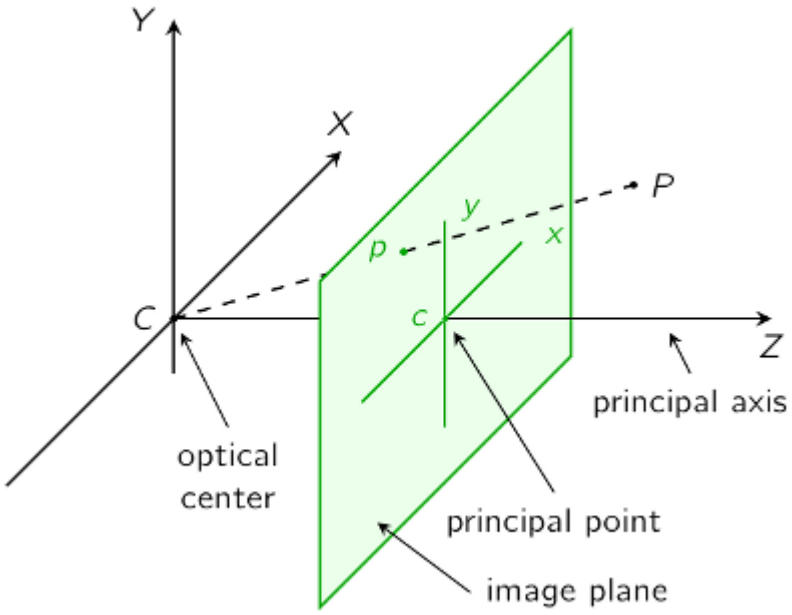


Figure 3.1: Pinhole camera model

(the position of the camera center with relative to the world coordinates). Here, R is a 3×3 rotation matrix and $\boldsymbol{T}$ is a vector in $\mathbb{R}^3$. These two quantities represent rotation and translation of the camera relative the world coordinate. To find all the parameters (K and g matrices) of the camera model we need to calibrate the camera [309]. In our case, the device is a 2.5D camera. In this kind of device the information acquired by the sensor is not the chromatic information (RGB) but an intensity value proportional to the distance of the point P (see Figure 3.1). These devices, however, still follow the pinhole camera model [114], but the camera calibration procedure is different [307], [121], and the final parameters are still the same as in the above equation.

We calibrate the Microsoft Kinect for Xbox [199] using the method in [121] and we use the calibration intrinsic data to simulate this camera in our Virtual Environment framework.

Apart from the intrinsic parameters in the pinhole model, the Virtual Environment also needs to account for other non-ideal behavior of the device. The geometric characteristics

of the camera are captured in the camera model, but we need to account for the electrical characteristics of the sensor. The sensor and the electrical components connected to it convert the light into electrical signals, and then into digital signals (gray level intensities). In this process, the signal is typically corrupted by noise, which in the case of a 2.5D device will result in distorted surfaces.

Some general methods can be used to de-noise the depth map, and some proved to be very useful. However, in our Virtual Environment, the goal is to simulate a real camera, using a model that can replicate the real camera behavior. We implement the method proposed by Nguyen et al. [206]. This method measures both lateral and axial noise distributions, as a function of both distance and angle of the Kinect to an observed surface. Using this procedure, we can simulate different scenarios, add noise to the final acquisition, and implement de-noising strategies able to reduce the effect of noise on the WBSA calculation.

3.2.3 Whole Body Surface Area from a Single View

The Body Surface Area is the 2D area of the external body skin. In our case, we are using the virtual subjects mesh as an approximation of the skin area. Common meshes are composed of vertices, faces, and edges. The faces can be regarded as 2D polygonal with the given vertices that constitute the surface of the 3D object. Figure 2.2 shows the wireframe representation of the body mesh used. An object acquired with a 3D scanner can have around 50000 faces. The computation of the total area of a mesh is nothing more than the sum of the 2D area of each face [174]. The area calculation can be done using common geometric formulae utilizing the edge lengths of each face. Unfortunately, there are some complications in this apparently simple operation. As mentioned, the human body can assume a large variety of poses, and it can assume different shapes changing the observation angle. In this situation, occlusions and surface curvature make the area calculation more complicated than usual.

The result of the ray cast method is a point cloud obtained by the intersection of the rays with the subject (Figures 3.12). The density of the point cloud depends on the resolution of the sensor and the distance from the camera (equation 3.1). From the point cloud, we need to reconstruct the mesh surface to be able to calculate the surface area. The literature on this topic is vast, especially in computer graphics. Traditional methods include marching cubes [185], Poisson surface reconstruction [143], greedy surface reconstruction [68]. All these methods present different reconstruction performance that varies depending on the surface complexity. Unfortunately, this step cannot be avoided, with the noise associated with it, since is fundamental in every 3D system that uses surface data. However, to be able to analyze the VBSA-WBSA without the reconstruction noise we need to calculate the mesh triangles area visible by the camera directly from the original mesh. We realized this approach keeping a list of the observed triangles. For each ray incident on the mesh surface, we store the triangle ID a list. At the end of the ray casting method we order, sort and eliminate multiple ID inputs, given by multiple rays intersecting the same triangle, and finally, we compute the areas of the triangles on the list. The sorting and elimination task is required since multiple rays can intercept the same triangle. This method gives us accurate results, but it can overestimate the real area. In fact, if a triangle is partially visible, this method will still compute the whole triangle area. However, since each body mesh is composed of roughly 28000 triangles, each triangle has a minimal contribution, and a portion of a triangle has an even smaller contribution.

With this method, we can study the relationship between the WBSA of a subject, and the VBSA, related to the camera position, without the additive noise coming from the surface reconstruction process. Interestingly, the developed framework is the ideal setting to test a new reconstruction algorithm since we can analyze different sources of errors that are camera position dependent. For example, under this framework, we can acquire the ground truth of the surface area directly from the original mesh, then we can calculate the distortion introduced

by the reconstruction algorithm, and the prediction error using the VBSA. All these processes using our framework can be treated separately, each with its additive noise model.

Surface Area Calculation

As an application of the presented framework, we analyze the relationship between WBSA and VBSA while varying the camera position. The initial surface area value for the whole mesh has been calculated from the MH plugin and stored with the subject measurements. After ray casting, we need to calculate the surface area from the visible part of the mesh (VBSA). The surface area algorithm is the same as the one used in MH. Given the edges $\mathbf{u}$ and $\mathbf{v}$ (see Figure 3.2) of a triangle, to obtain the surface area, we use the standard relation:

$$A = \frac{1}{2}|\mathbf{u} \times \mathbf{v}| \tag{3.3}$$

Where $\times$ denotes the cross product between the two vectors $\mathbf{u}$ and $\mathbf{v}$, and $|\ |$ denotes the magnitude of the cross product. The magnitude of the cross product is the area of the parallelogram whose edges have length $\mathbf{u}$ and $\mathbf{v}$ (see Figure 3.2). This is twice the area of the triangle whose edges are $\mathbf{u}$ and $\mathbf{v}$. The result of this operation is the surface area of a single triangle. Our initial MH mesh is composed of about 28000 faces, but given the simplicity of this operation, it can be done almost in real time.

3.2.4 WBSA Prediction

Given the VBSA value, we want to predict the WBSA for each position of the camera. We expect a behavior somehow linear when the camera is front to the subject, but will rapidly diverge when overlapping areas and fat subjects are examined. In this section, we're going to introduce the statistical model used for the prediction.

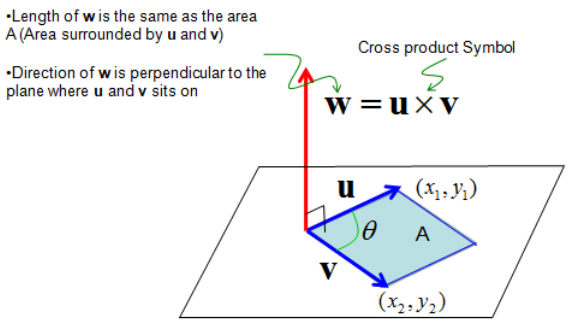


Figure 3.2: Mesh surface calculation

WBSA-VBSA formulation

We use a polar coordinate system (Figure 3.3), where the subject is at the origin, and the camera is free to move on a sphere with the same center pointing the origin, and radius the distance from the subject.

Given the VBSA observations from a given view, say at (θ_0, ϕ_0) at a fixed $distance = r$:

$$\mathbf{VBSA}(\theta_\mathbf{0}, \phi_\mathbf{0}) = (\mathrm{VBSA}_1(\theta_0, \phi_0), \mathrm{VBSA}_2(\theta_0, \phi_0), \mathrm{VBSA}_3(\theta_0, \phi_0), \dots)^T \tag{3.4}$$

we want to infer the WBSA from a single observation at angle (θ_0, ϕ_0) : $\mathrm{WBSA} = f(\mathrm{VBSA}(\theta_0, \phi_0))$. Due to the symmetric nature of the human body, the visible portion is close to half of the total body area. Although, in [304], where the body is scanned by parts and subsequently fused together, the frontal part of the body account for more than $50\%(52\%)$ of the WBSA. This means that the front and the back contribute in slightly different proportions. In general, for a given angle, the surface visible is directly proportional to the body dimension, then to the WBSA. VBSA of that view. The approach is as follows. We process the bodies generated in the Virtual NHANES dataset and Virtual Random dataset with the Virtual Environment, positioning the camera at different orientations. The set of all positions of the camera span a solid angle covering almost all the possible camera views of the body. The solid angle was chosen within

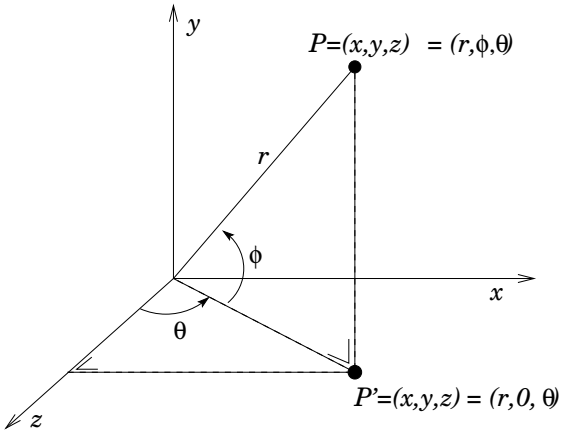


Figure 3.3: Polar coordinate system.

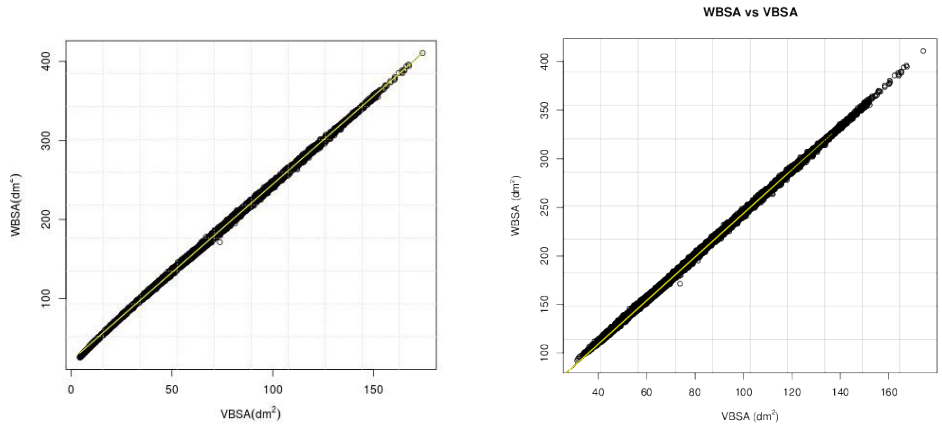
$-90 \leq \theta \leq 90$ and $-90 \leq \phi \leq 90$. Since the virtual datasets are composed of symmetric subjects, we limit the azimuth angle on the left side of the subject covering the angles from the front left side to the back left side (see Figure 3.3). We limit the body pose to the default pose in Figure 2.2 and maintain a constant distance between the subjects and the camera. This distance has been found empirically by considering the tallest subject in the dataset. However, using different distances, we did not see any drastic difference in performance when the range comprises in the $3.5 - 4.5$ meters.

Statistical analysis

To find the relation between WBSA and VBSA, we need to assume the statistical model to be used for the inference. From a first plot of VBSA vs. WBSA (Figure 3.4), we can see that a linear regression model can potentially obtain good results. With this assumption the vectors $vbsa(\theta, \phi)$ and $I(\theta, \phi)$ in equation 3.5 are the unknowns of our system:

$$WBSA = \alpha(\theta, \phi)VBSA(\theta, \phi) + I(\theta, \phi) \tag{3.5}$$

where I is the intercept and a is the view area linear coefficient. To avoid overfitting in the prediction, we use a k-fold cross validation with k=10. We repeat the fitting for different par-



(a) Virtual Random VBSA-WBSA relation for $\theta = 0°\phi = 0°$

(b) Virtual NHANES VBSA-WBSA relation for $\theta = 0°\phi = 0°$

Figure 3.4: WBSA-VBSA relation.

titions of the dataset: males, females, kids, adults, small stature ($s \leq 140cm$), normal stature ($s = 140 - 200cm$), big stature ($s > 200cm$). Another interesting analysis is the use of some measurements in the prediction. As we can see from the correlation matrix in Figure 3.5, some measurements are highly correlated with the WBSA ($\rho > 0.9$). But since our actual system permits to acquire the stature with good accuracy, we conduct this analysis with the intent to show some possible gain in the use of body measurements. Using the stature, the inference takes the form of a multi-linear regression:

$$WBSA = \alpha(\theta, \phi)VBSA(\theta, \phi) + \beta(theta, \phi)Stature + I(\theta, \phi) \tag{3.6}$$

Since we work with a calibrated camera, and due to the morphology of the human body, we can retrieve the stature almost from every angle, and it's independent of the view angle. It's also worth mentioning that the stature is the most accessible measure to acquire and most reliable.

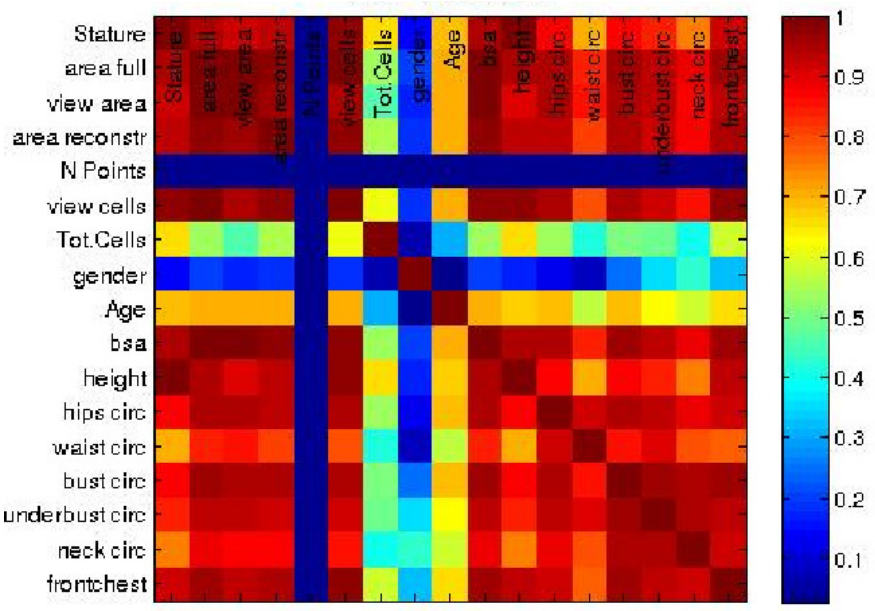


Figure 3.5: Correlation Matrix.

Theoretically, having a system able to capture other measurements will possibly lead to a better prediction.

3.3 Results

In this section we present the result for the WBSA prediction.

3.3.1 WBSA Prediction

A first interesting analysis is the correlation of the available quantities (WBSA, VBSA and body measurements). Figure 3.5 shows the correlation matrix for the analyzed quantities for all subjects in the Virtual Random dataset for $(\theta = 0°, \phi = 0°)$. We use the Spearman's ρ as our statistic. The WBSA is strongly correlated with the VBSA ($\rho = 0.9992$). The WBSA is also strongly correlated ($\rho > 0.9$) with the following quantities: stature, hip circumference,

frontal chest. Other measures that have high correlation ($\rho > 0.8$) with the WBSA include waist circumference, bust circumference, underbust circumference and neck circumference. The correlation between WBSA and stature is trivial since the stature is one of the parameters directly connected with the body surface area (all the WBSA formulae are based on stature and weight). The correlation is an interesting indicator since it can help us to determine which parameters can give a better prediction of the WBSA.

3.3.2 Linear Regression Analysis

Figure 3.4 shows the scatter plot using VBSA and WBSA in abscissa and ordinate respectively. Each subject is represented by a point of coordinate (VBSA, WBSA). The relation is linear. Thus we analyze the performances of a linear regression model. We use the R regression DAAG Tool model [232] to find the unknowns of the model. We use a k-fold (k=10) cross-validation to calculate the prediction error for the models defined in Equation 3.5 and Equation 3.6. We repeat the experiments using different partitions of the data (all subjects, males, females, adults, kids, small stature, normal stature, big stature).

Figures 3.6-3.8 show the results of this prediction. Each row corresponds to an orientation of the camera concerning the subject. For each subject, we analyzed the angles $\theta = 0°, 30°, 45°, 60°, 90°, 120°, 135°, 150°, 180°$ for the azimuth, and $\phi = 0°, \pm 30°, \pm 45°, \pm 60°, \pm 90°$ for the elevation. For $\phi = \pm 90°$ we only had $\theta = 0°$ since changing the azimuth angle will not affect the view area. In the tables we report some statistical indicators needed to evaluate the fit as t-value and standard deviation and different prediction errors given by R: residual standard error, cross-validation root mean square error, cross-validation mean square prediction error, cross-validation mean absolute prediction error.

We calculated these quantities for $(\theta = 0°, \phi = 0°)$ position of the camera; we'll see later that for some view of the camera, the model will diverge from the linear model, and will see if

a linear model still can be used with good performances.

As we can see from Figure 3.4 the relation is linear and can be easily fitted. The values for Multiple R^2 and adjusted R^2 are very high (0.9975) that means that the residuals are closer to the linear model (low variance), but doesn't tell much about the best fit. From Figure 3.4 we can see that the residuals are close to the median value, and hence the R^2 value is high, but in this situation, we cannot use this value for establishing which prediction is better. A more useful statistic is the standard error (SE) of the residuals. Also indicative is the distribution of the residual given by min max 1st and 3rd quantile.

3.3.3 Impact of Azimuth and Elevation on Computed WBSA

As expected, the camera orientation (as captured by the azimuth θ and elevation ϕ) has a significant impact on the computed WBSA, see Figures 3.6 and 3.7. Figures 3.6a and 3.7a show the variation of the VBSA linear coefficient (α) and the regression RMSE (Root Mean Square Error) as the azimuth angle change for different elevation angles. For azimuth angle $\theta = 0°$, the subjects appear with the maximum area (VBSA) of the body facing the camera, and the linear coefficient (α) is at the minimum value. As the azimuth angle increases to $\theta = 90°$, the VBSA decreases, and the linear coefficient (α) increases. At $\theta = 90°$, the area facing the camera is at the minimum since it's the angle where the camera can see only one side of the body. As the azimuth angle goes from $90°$ to $180°$, the body shape is similar to the frontal part, but due to the body pose, the occluded areas make the difference: the VBSA increases, since the area facing the camera increases and the coefficient α decreases. For the cross-validation error, (Figure 3.6a left), the RMSE has a singular behavior. For all the azimuth angles $\theta \leq 90°$ the error increase, but it reaches the maximum at $\theta = 60°$ and not for $\theta = 90°$ as predicted. This unexpected behavior is confirmed for the elevation angles $\phi = 30°, 45°, 60°$. Intuitively the azimuth angle with the lowest accuracy should be $\theta = 90°$ because the body presents less area to the camera.

Instead, this is true for the angle $\theta = 60°$. The explanation for this behavior is that for this angle the body presents more overlapped areas. For the reciprocal angle, $\theta = 150°$ it doesn't happen, because the arms slightly bent upwards it does not affect that view and make the prediction for $\theta = 60°$ the one with the worst accuracy. This behavior is not happening for every elevation angle. At $\phi = 0°$ the lowest accuracy is at $\theta = 90°$ because with the camera aligned with the body center the maximum overlap is observed at $\theta = 90°$, but, as the elevation increase, less and less area faces the camera thus more artifacts can appear.

Varying the elevation angle, just as changing the azimuth, less area is visible by the camera. But, differently, from the azimuth case, this behavior is not linear and smooth as described before. In fact, observing the plots in Figure 3.9 we can see that the VBSA coefficient does not always increase linearly. See the results at $\theta = 30°, 45°, 60°$. For these angles, the overlapped areas, due to the left arm and leg, make the relation to divert from being linear. The intercept I has a maximum at approximately $\theta = 60°$. The intercept is associated with the bias of the prediction, an offset to add to the linear increase of the VBSA. The WBSA and VBSA are linearly correlated. For $\theta = 60°$, there are many occlusions, and the legs are overlapped so the other arm that remains almost entirely occluded. Thus we can see that the relation diverts from being linear.

To evaluate the prediction we used a k-fold cross validation setup. We randomly compose the folds using the function from the CVTOOL package in R [231]. We measure BSA prediction performance regarding prediction errors: root means square error (RMSE) under k-fold cross-validation, denoted as CV RMSE in the tables, but also CV MSPE (mean squared prediction error) and CV MAPE (mean absolute prediction error). Since the WBSA and VBSA are calculated from a mesh with the base unit in decimeter (dm), the WBSA is in decimeter squared (dm^2). The cross-validation root means square error (CV RMSE) is in decimeter squares too, while the cross-validation mean absolute prediction error (CV MAPE) is in percentage ($\%$).

CV RMSE varies from a maximum value of $5.5\ dm^2$ when the camera is at the unfavorable position of $90°$ in elevation, to a minimum of $0.71\ dm^2$. These can be considered relative to the average WBSA, for the virtual dataset, the average is $167\ dm^2$, and the relative error is 0.6%. The behavior of the CV RMSE is not straightforward, but it seems connected with the occluded areas and consequently with the position of the camera. In all the predictions CV RMSE was higher for the positions of the camera in front of the subject (azimuth 0-90) and has a high value at around $\theta = 60°$. The highest value, however, is for the elevation of $90°$ where the camera can see only the footprint of the body. In this case, the stature component is missing from the image, but the prediction can still get a decent estimate for obese subjects.

3.3.4 Regression with Stature

Since the WBSA is high correlated with the stature (see Figure 3.5), we expect improved results by including the stature in the prediction model (Equation 3.6). Figures 3.6b, 3.7b and 3.8 show the impact of including stature in the model. In general, we obtain a lower prediction error for most analyzed angles, as can be seen in Figure 3.8 thought the improvement might not be as significant in some cases, e.g., as azimuth angle θ moves away from $90°$. However, it's not always possible to acquire the stature with accurate precision; this is not the case in a physician's clinic, where the controlled environment, always permits the detection. However, in a more unconstrained environment, we should consider the stature detection error and its influence on the WBSA estimation. Out of the scope of this work.

3.3.5 Regression with Grouping

We investigate the performances of the system for different specified human categories. We grouped our virtual subjects into five different classes (males/females, adults/kids, small/nor-

mal/big stature), and we use these partitions to learn the linear system. Then using 10-fold cross validation, we compute the prediction error. Figures 3.6, 3.7 and 3.8 show the VBSA coefficient α, the RMSE error and a comparison of the errors for cases with and without grouping. Grouping has a different effect on the prediction error. Surprisingly, grouping did not always lead to an improvement. In fact, the errors for adults seems larger than the error obtained for all subjects. Instead, we have some improvement for kids, small stature, and normal stature. Using the stature in the group models results in a significant improvement for most groups (see Figure 3.6,3.7).

3.4 Discussion

In this work, we presented an integrated computer vision framework able to infer the relationship between the Whole Body Surface Area (WBSA) and the View Body Surface Area (VBSA) for a given viewpoint of the subject. In this section, we discuss some observations from the obtained results. Figures 3.6a,3.7a,3.8 show the WBSA prediction errors for different experimental settings using the two datasets. Other plots that show the WBSA error behavior can be found in Supplementary Material. In all the WBSA prediction plots we can see a logically natural pattern: the error remains low for azimuth angles between $0 - 45°$ and $135 - 180°$, but higher for $60 - 150°$. From this behavior, we discuss some interesting observations, some of which are not so apparent. Some of these effect are very difficult to observe.

3.4.1 Frontal VBSA Vs Rear VBSA

An interesting observation is the difference in behavior of the VBSA (and hence computed WBSA) when the subject is viewed from the front or from the back. As reported in [83] and [304], the front accounts for more than 50% of the total WBSA. As we can see from the RMSE

(Figures 3.6a, 3.6b), the errors from the rear part are always inferior relative to those from the corresponding angles from the front. There could be several explanations. Since the hands are bent slightly upfront, the occlusions are greater seeing the mesh from the front. Moreover, the frontal part of the human body and consequently the mesh has many more curved surfaces on the front, and hence more challenging to model. This effect can be noted by comparing the RMSE error from males and females. Despite some irregular behavior (males have higher RMSE at $\theta = 90°$, than at $\theta = 60°$) the average RMSE for the males is lower, for the frontal angles ($0° \leq \theta \leq 90°$), than females.

3.4.2 Non-Linearity in the WBSA-VBSA Relationship

Figures 3.6, 3.7 and 3.8 show the RMSE error increase as the azimuth angle, approaches $90°$. The same behavior can be seen as the elevation angle ϕ approach $\pm 90°$. Intuitively, for these angles, the VBSA can hardly infer the WBSA of the subject (Figure 3.11). In these situations, there are many overlapped areas other than the usual occlusions (feet, armpit, crotch, etc.). Since we obtain a higher error in a linear prediction, that means that the relation VBSA-WBSA diverges from linearity. Figure 3.10 shows the VBSA-WBSA relation for the two datasets at $\theta = 60° \quad \phi = 60°$. Observe that for these angles the homoscedasticity condition (constant variance) fails. People with small WBSA have small variance, instead, big people have very large variance. In this situation, a linear model can still be used, assuming that we accept a slight decrease in performance, for subjects with small WBSA, but it will strongly impact the performance for high WBSA subjects. Figure 3.8 shows the performance of the linear model for different categories. For high and very high WBSA values, a different approach (non-linear) has to be considered.

3.4.3 Evaluating WBSA Measurements

One problem in every WBSA study is the availability of data. In this kind of studies, real people have to be measured to constitute the ground truth. Our method instead, uses the ground truth of virtual subjects. We generate these subjects making sure that they are very close to real people. Although, we cannot compare our method with the familiar formulae. In fact, our lab is not equipped with wraps and mold to compute the WBSA. Also, the 3D acquisition through a cumbersome scanner doesn't guarantee the correctness of the method.

3.4.4 Reconstruction

As explained in Section 3.2.3, the WBSA retrieval is based on the triangular mesh area calculation. The subjects in the datasets are represented as a mesh with the WBSA calculated from the MH plugin. However, the ray cast result is a point cloud as shown in Figures 3.11, 3.12. A fundamental step to retrieve the VBSA is surface reconstruction. Since surface reconstruction is a hot topic in computer graphics and is beyond our goals in this work, we decided to study its impact with just one known algorithm: the greedy surface reconstruction algorithm [68]. In this experiment, we use the default setting that should give good results.

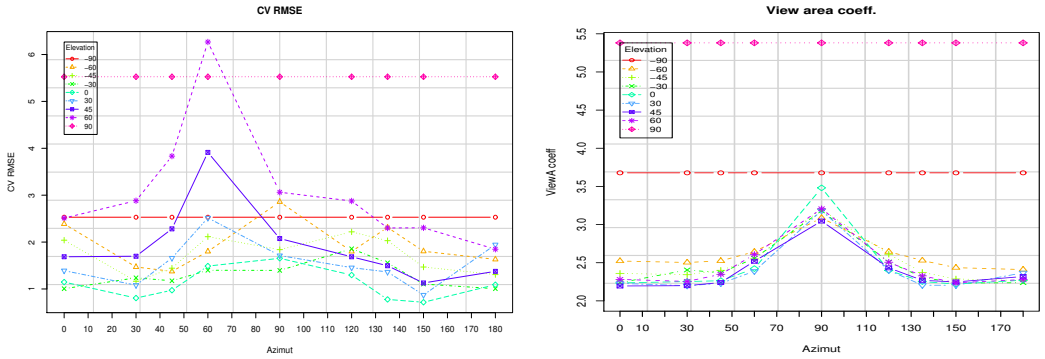
Before applying the reconstruction algorithm, an intermediate step is the points normal calculation. For this we used an algorithm based on integral images [129] implemented on the PCL library. We estimated a reconstruction time of $\approx 0.5s$ in average for one subject (subjects with more overlapped areas slow down the algorithm). Figure 3.13a shows the cross-validation error for the linear regression using VBSA computed from the reconstructed surface. We observe that, at the indicated elevation angle $\phi = 0°$, the errors are still generally lower for the azimuth angle $0° \leq \theta \leq 45°$ and $135° \leq \theta \leq 180°$ (less than about 6% for MAPE, and less than about 5.0 for RMSE).

These results show the reconstruction error as an additive noise on the VBSA. This noise is composed of two main components. The first is the error due to the points normal. In fact, errors in the normal direction will impact the subsequent surface reconstruction. Unfortunately, due to the very complex nature of the human body and the additional complexity due by the perspective view in the ray cast operation, computing the normals is not that easy. Missing neighboring points, surfaces with weird angles due to the non-rigid nature of the body make this operation complex and prone to errors. Figure 3.12 shows the result of the normals calculation. The surface reconstruction operation is the second source of the noise. This basic operation is responsible for transforming a raw or basic representation of the subject (i.e., a cloud of data points) into a closed manifold mesh. One of the main challenges to surface reconstruction algorithms is hole filling. A hole in the mesh structure is possibly caused by gaps in the mesh structure, which if left untouched would result in a surface with numerous jagged boundaries. This phenomenon is the main source of error in the surface area calculation. In fact, since we just compute the areas of the single triangles, erroneous reconstruction will create unwilling boundaries that increase the calculated surface area. This behavior has been observed during the software setup. To correct this effect, smoothing with least mean square and sampling with voxels algorithm stages have been added before the final reconstruction.

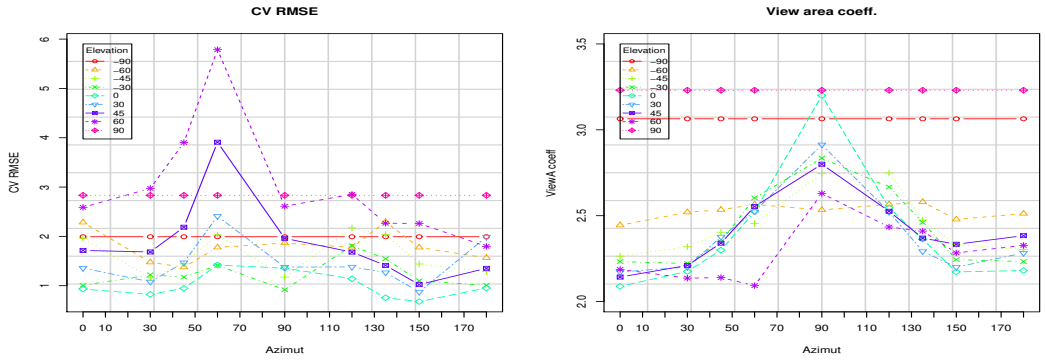
Usually, every surface reconstruction algorithm is tested using SSD or similar measures. Unfortunately, since the surface area computation is based on triangle area computation, the usual measures don't always consider the reconstructed topology of the final mesh. These distortions in the topology can drastically degrade the surface area calculation.

All these methods are accurate and can produce a reliable surface. However, they need a significant amount of time to reconstruct the partial surface of the subject. Simulating a 2.5D device, gives us what is called an organized point cloud (the x,y,z points are organized in a matrix fashion like the pixels of an image), and we can use faster and simpler methods for the

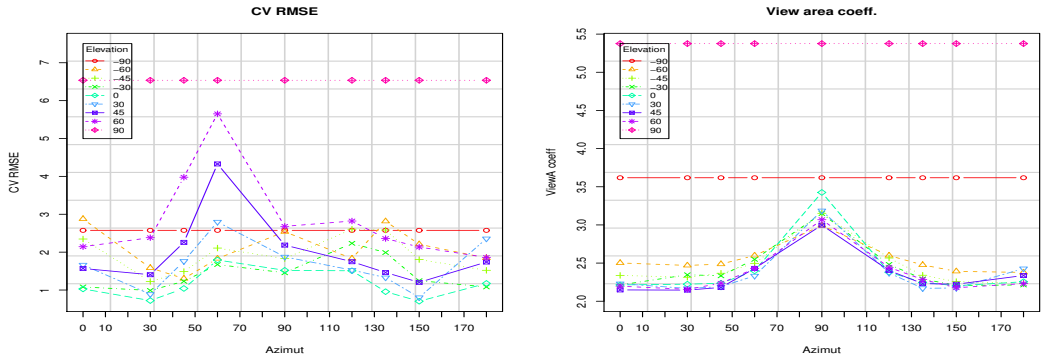
reconstruction, for example [128].



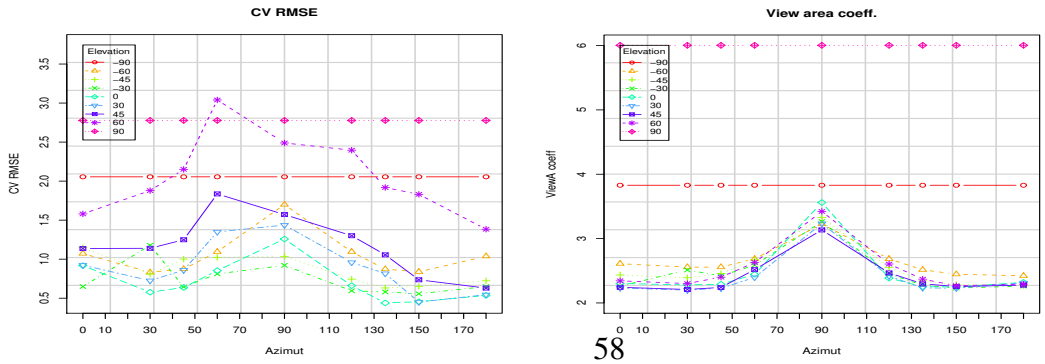
(a) All subjects: RMSE error (left), VBSA coeff. α (right)



(b) All subjects with stature: RMSE error (left), VBSA coeff. α (right)

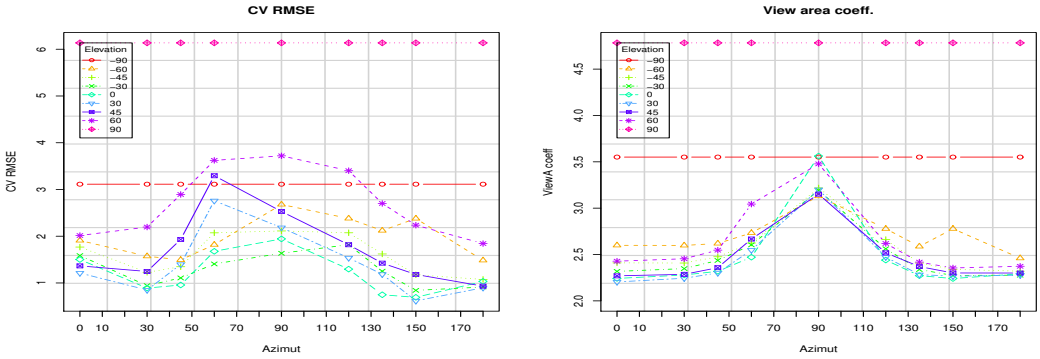


(c) Adults: RMSE error (left), VBSA coeff. α (right)

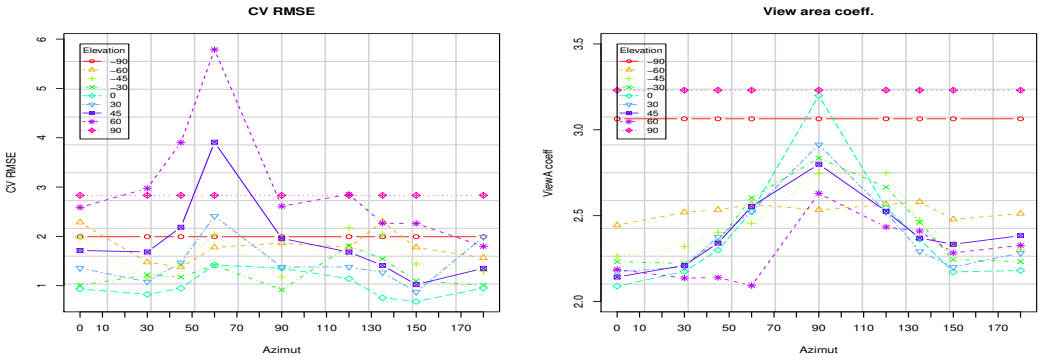


(d) Kids: RMSE error (left), VBSA coeff. α (right)

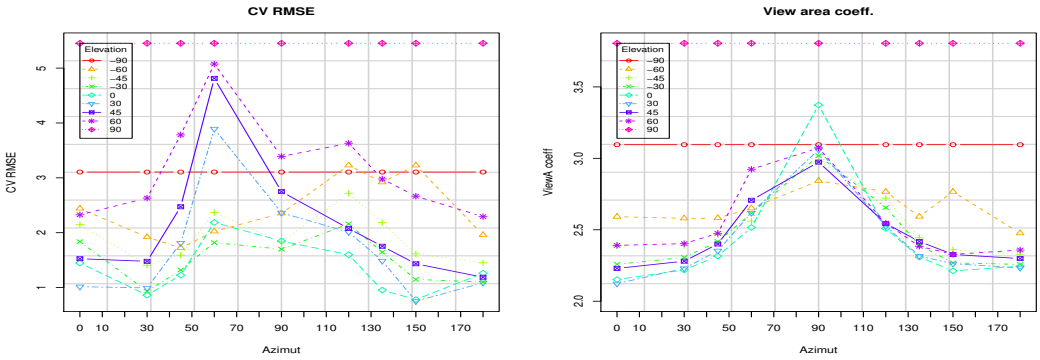
Figure 3.6: Virtual Random dataset.



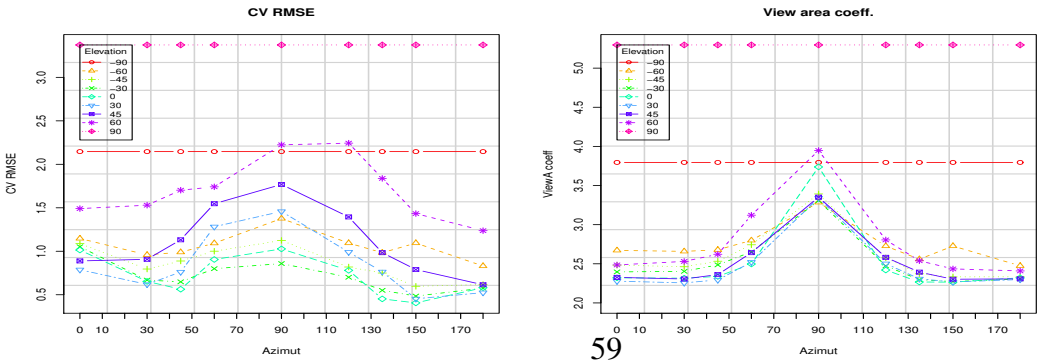
(a) All subjects: RMSE error (left), VBSA coeff. α (right)



(b) All subjects with stature: RMSE error (left), VBSA coeff. α (right)

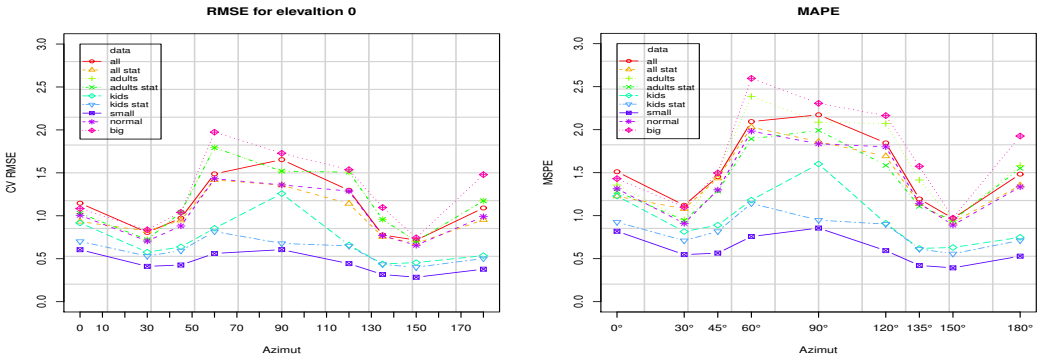


(c) Adults: RMSE error (left), VBSA coeff. α (right)

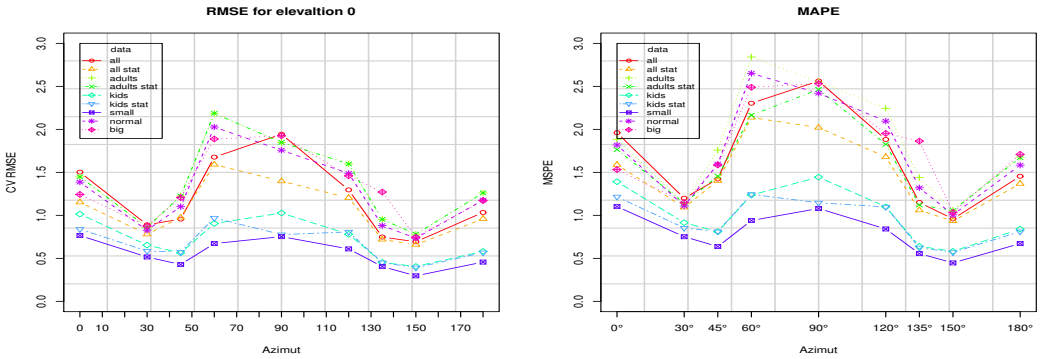


(d) Kids: RMSE error (left), VBSA coeff. α (right)

Figure 3.7: Virtual NHANES dataset.

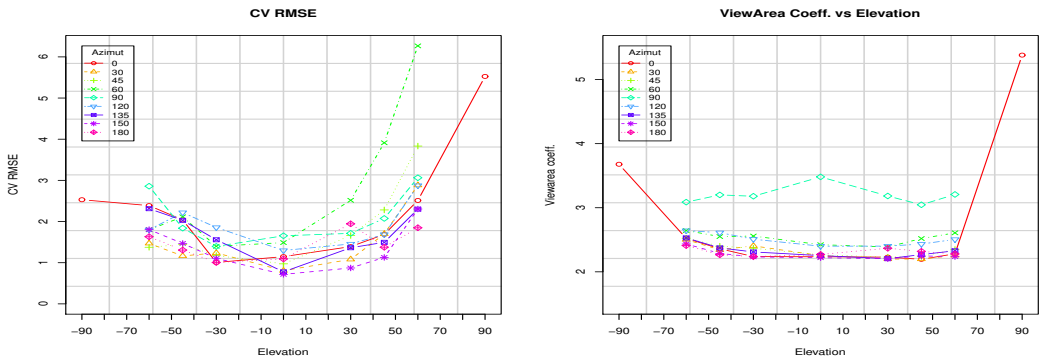


(a) Virtual Random alla subjects.

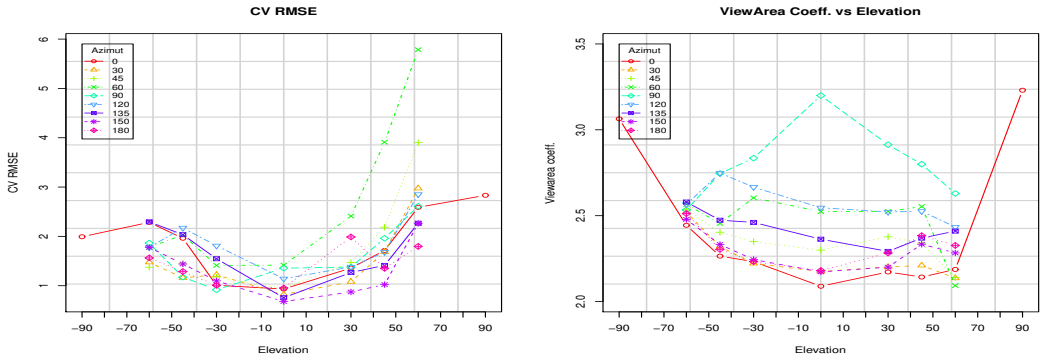


(b) Virtual NHANES all subjects.

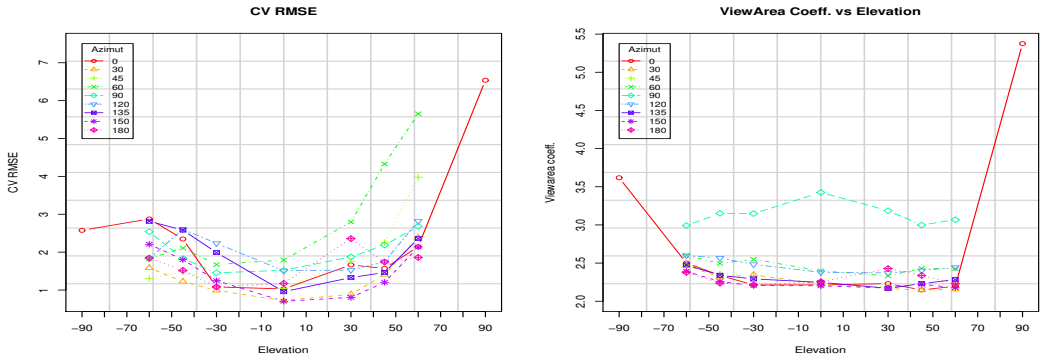
Figure 3.8: WBSA prediction errors at elevation angle $\phi = 0°$.



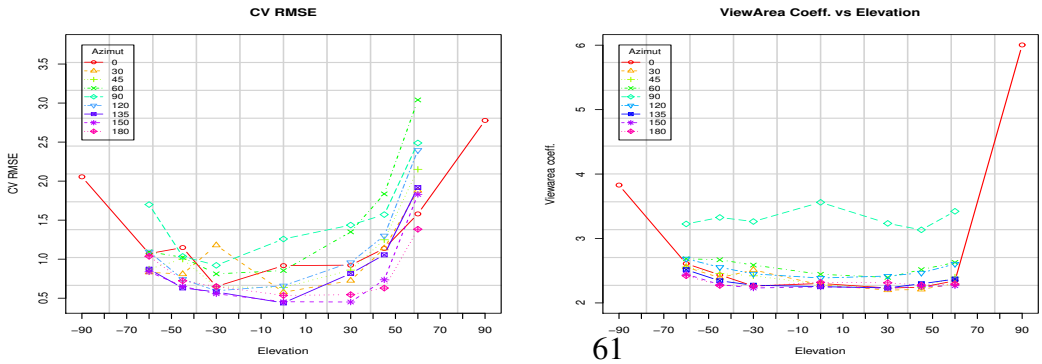
(a) All: RMSE error (left), VBSA coeff. α (right)



(b) All stature: RMSE error (left), VBSA coeff. α (right)

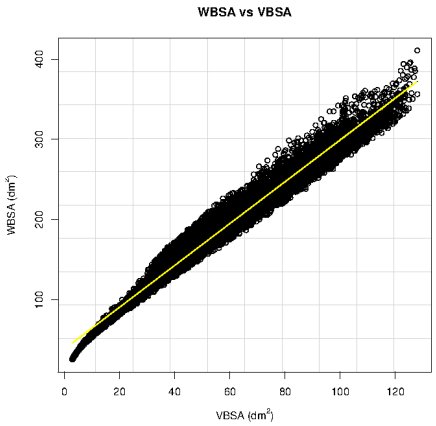


(c) Adults: RMSE error (left), VBSA coeff. α (right)

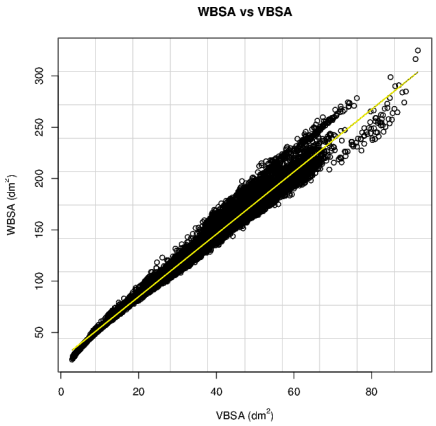


(d) kids: RMSE error (left), VBSA coeff. α (right)

Figure 3.9: Virtual Random dataset. Impact of camera orientation (Azimuth and Elevation) on the VBSA prediction.



(a) Virtual Random VBSA-WBSA relation for $\theta = 60°\phi = 60°$



(b) Virtual NHANES VBSA-WBSA relation for $\theta = 60°\phi = 60°$

Figure 3.10: Relationship between VBSA and WBSA.

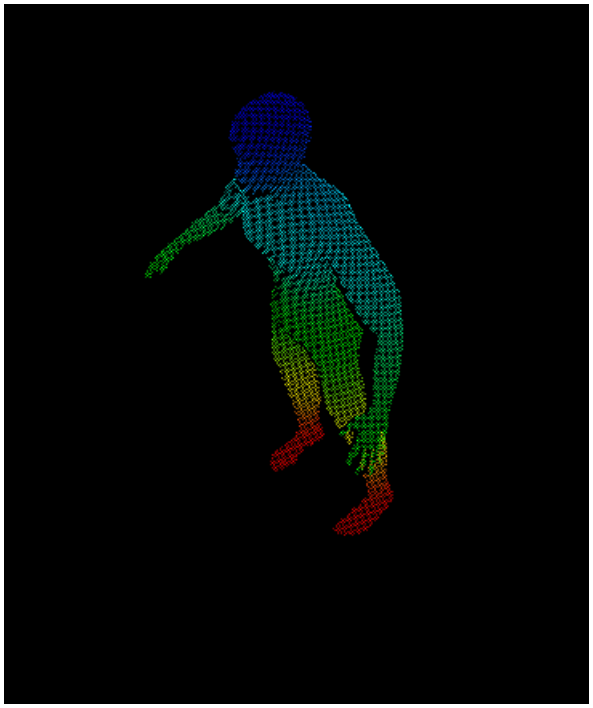


Figure 3.11: Point clouds from raycast. Subject 8 from Virtual NHANES dataset at $\theta = 60°$ $\phi = 60°$.

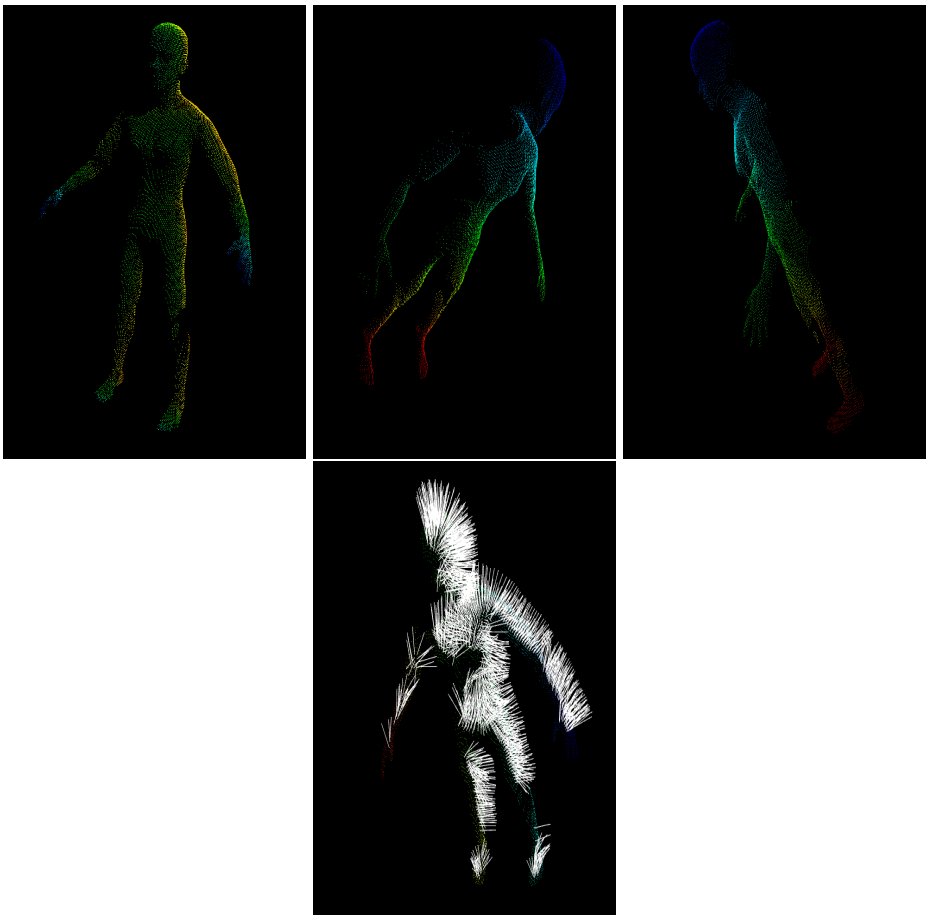
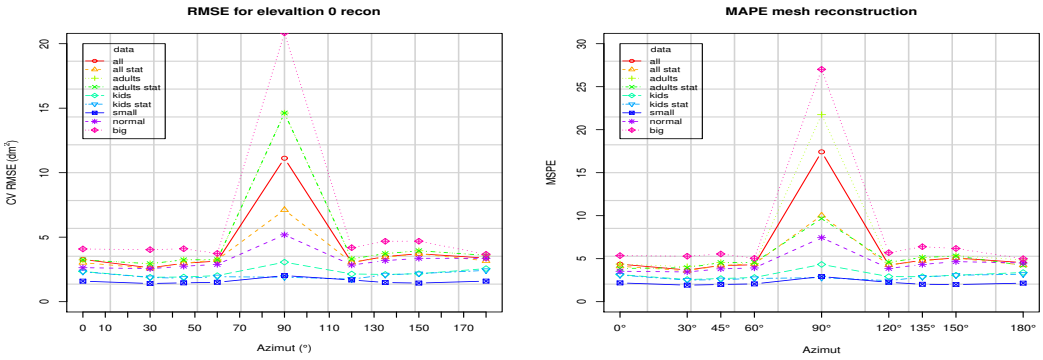
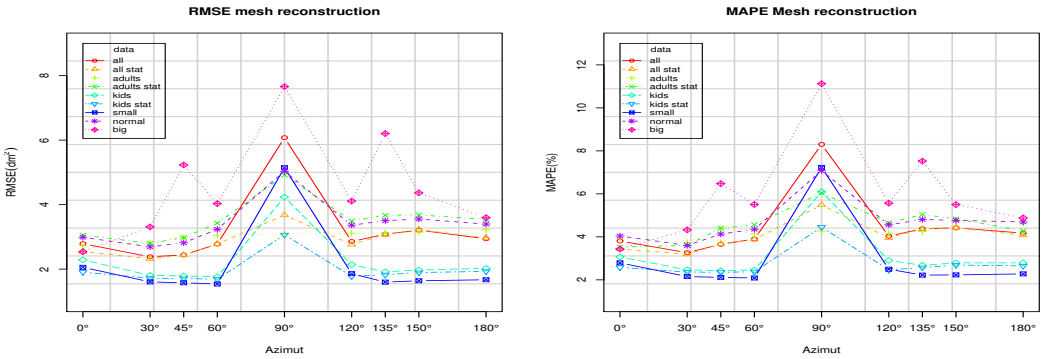


Figure 3.12: Point cloud results from Virtual Environment. Subject 8 from Virtual NHANES dataset at $\theta = 60^\circ$ $\phi = 60^\circ$ seen by different angles. From these shots, it is possible to see the missing parts of the body as result of raycasting operation with the camera at the above angle.



(a) Virtual Random dataset.



(b) Virtual NHANES dataset.

Figure 3.13: Mesh reconstruction results at elevation angle $\phi = 0°$.

Chapter 4

3D Body Shape Analysis

In the previous chapters we introduced a new dataset, then we used it to find a useful relationship between the whole body surface area and the visible part of the body surface. In this chapter, we start a more detailed study on body shape and its description. After a brief introduction of related topics in computer vision and geometry processing, we propose techniques to classify body shapes in terms of their Body Fat Percentage (BFP), a label contained in the introduced dataset. The analyses contained in this chapter are based on the entire mesh structure, and some newly introduced operators that are intrinsic to the mesh surface. As we'll see, using **Spectral Geometry** techniques in this setup constitute a novelty with enormous potentials.

4.1 Shape Analysis in Computer Vision

Over the past 40 years, a vast collection of work has been devoted to the definition and analysis of the shape, and shape spaces, as mathematical objects, and to their applications to various domains in computer graphics and design, computer vision and medical imaging. In computer vision and medical imaging an important scientific field has started, initiated by U. Grenander

and M. Miller, called **computational anatomy** [109][108][110]. One of the primary goals of **computational anatomy** is to analyze diseases via their anatomical effects on the shape of the organs. Shape analysis has demonstrated itself as a compelling approach to characterize brain degeneration resulting from neuro-cognitive impairment like Alzheimer's or Huntington's diseases and has contributed to a deeper understanding of disease mechanisms at early stages [173].

Whether represented as a curve, or a surface, or as an image, a shape requires an infinite number of parameters to be mathematically defined. It is an infinite-dimensional object, and studying shape spaces requires mathematical tools involving infinite-dimensional spaces (functional analysis) or manifolds (global analysis). Some examples are reviewed in the survey paper from Bauer et al. [18]

In the context of **pattern theory** [107], a shape is represented as a deformation of another (fixed) shape, called template. The deformable template paradigm is rooted in the work of D'Arcy-Thompson in his celebrated treatise (On Growth and Form) [283], and developed in Grenander's theory. Even if **pattern theory** can be more general, recent models of deformable templates in shape analysis focus on deformations represented by diffeomorphisms acting on landmarks, curves, surfaces or other structures that can describe shapes [272].

In the last two decades, the attention has been mainly on feature descriptors. In fact, feature descriptors play a crucial role in a wide range of geometry analysis and processing applications, including shape correspondence, retrieval, and segmentation. For 2D images, well known descriptors like SIFT [186], HOG [203], MSER [194], and shape contexts [21]. Early works in geometry processing such as spin images [138], shape distributions [210], and integral volume descriptors [192] were based on extrinsic structures that are invariant under Euclidean transformations.

The next generation of feature descriptors introduced *intrinsic* structures, such as geodesic

distances [85] that are preserved under isometric deformation and entirely invariant for the embedding. The geodesic distance, however, suffers from strong sensitivity to topological noise, which limits its usefulness in real applications. A different branch of shape analysis has been developed with the use of **harmonic methods** on *surface embedded in 3D manifolds*. The origin of this methodology can be traced to the original article "Can One Hear the Shape of a Drum?" [140] published by Mark Kac in 1966.

The frequencies at which a drumhead can vibrate depends on its shape. These frequencies are the eigenvalues of the Laplacian in the space. A central question "can the shape be predicted if the frequencies are known?" known as inverse problem. Is it possible for two different shapes to yield the same set of frequencies? The answer came in 1992 when Gordon, Webb, and Wolpert constructed a pair of regions in the plane that have different shapes but identical eigenvalues [105]. So, the answer to Kac's question is: for many shapes, one cannot hear the shape of the drum completely. However, some information can be inferred.

Unfortunately, **the spectrum does not completely determine the shape of the underlying manifold, even though geometrical data is contained in the eigenvalues.** Manifolds with identical spectra will be called *isospectral* manifolds. Although the **spectral analysis** cannot give a unique solution to **isospectral** objects, it can provide good properties that we list in Appendix B.1.1. In the next section, we discuss how the spectral content can be used in geometry processing.

Other techniques used for human shape analysis comprise the silhouette analysis [113],[120]. Despite the good results, these methods suffer from the single view perspective of the body, very difficult to solve without heavy constrain the problem or with important prior information (3D).

4.2 Spectral Analysis

In response to the question "can one hear the shape of a drum?", it is possible to hear the following information from the spectrum:

- It has been shown that if two compact Riemannian manifolds M and $\tilde{M}$ are isospectral, then $dim(M) = dim(\tilde{M})$ and (Riemannian) $volume(M) = volume(\tilde{M})$. Hence, the spectrum determines the dimension and the volume of a Riemannian manifold. McKean and Singer [196] showed the equality of the respective curvature integrals for the scalar curvature K (e.g., the Gauss curvature in case of a surface) for **isospectral** manifolds ($\int_M k = \int_{\tilde{(M)}} \tilde{K}$).

- In the case of a compact d-dimensional manifold M with a compact (d-1)-dimensional boundary B in addition to the previous results, the (Riemannian) volume of the boundary B can be heard [196]. However, to obtain the curvature integral of M and the integrated mean curvature ($\int_B J$) the spectrum of the double of M is needed.

- In the cases of a closed surface ($dim = 2$ human body mesh) and of a planar domain with a smooth boundary, McKean and Singer [196] deduced the possibility to hear the Euler characteristic from the spectrum. Thus, Kacs conjecture of hearing the number of holes in the case of a planar region M with smooth boundary B can be obtained. For surfaces with smooth boundary, the Euler characteristic and the geodesic curvature integral of the boundary curve can be obtained from spectral data as well, if one additionally employs the spectrum of the surface double.

Spectral Analysis has given rise to a long stream of works, especially in the areas of shape retrieval. Coifman et al. [69] introduced invariant metrics known as diffusion distances, which correspond to the L_2-norm difference of energy distribution between two points initialized with

unit impulse functions after a given time. The diffusion distance is more robust to topological noise than the geodesic distance. Subsequent works are based on the eigenvalues and eigenfunctions of the Laplace-Beltrami Operator (LBO) [118]. Since the result of this decomposition has a beautiful physical interpretation: the square roots of the eigenvalues $\sqrt{\lambda_i}$ are the eigenfrequencies of the membrane, and $\psi_i x_p$ are the corresponding amplitudes at x_p. In particular, the second eigenvalue corresponds to the sound we hear the best. The LBO [118] constitutes the Swiss-knife for all the works in **geometry processing**. For a more detailed treatise of the LBO, operator see the Appendix B.1 and the most recent book on **harmonic analysis** [118]. We briefly describe a few recent works on this topic below.

ShapeDNA, Reuter et al. [241]: Lévy [164] showed that the eigenfunctions of the Laplace-Beltrami operator could be well adapted to the geometry and the topology of an object.
Reuter et al. [241] adopted the eigenvalues of the LBO to construct a global shape descriptor, called ShapeDNA. At the heart of this method is the assumption that the Laplace-Beltrami spectra can be thought as the fingerprints for surfaces and solids. Since the spectrum is **isometry invariant**, it is independent of the objects representation including parametrization and spatial position. Additionally, the eigenvalues can be normalized so that uniform scaling factors for the geometric objects can be obtained easily. Therefore, checking if two objects are isometric needs no prior alignment (registration/localization) of the objects but only a comparison of their spectra. However, two non-isometric but isospectral solids that cannot be distinguished by the spectra of their bodies and present evidence that the spectra of their boundary shells can tell them apart.
ShapeDNA can be used (like DNA-tests) to identify objects in practical applications. As in real life, the DNA does not completely characterize a subject. Identical twins exist with different shape but the same ShapeDNA. Even though these twins are shaped differently, they still have

quite a few familiar geometric properties (precisely those properties that are determined by the spectrum).

Heat Kernel Signature, Sun et al. [276]: An important work on feature descriptors for deformable shape analysis is the Heat Kernel Signature (HKS). HKS is based on the equation of heat diffusion over a surface. Given the well-known heat equation:

$$\left(\Delta + \frac{\partial}{\partial t}\right) u(x;t) = 0 \qquad u(x;0) = u_o(x), \quad u(\partial\Omega) = \ldots \text{(boundary conditions)} \tag{4.1}$$

Where $u : \Omega \in \mathbb{R}^m \times \mathbb{R}^+ \rightarrow \mathbb{R}$ is the heat distribution at point $x \in \Omega$ at time $t > 0$, $u_0(x)$ is the heat distribution on the surface at time $t = 0$. Δ is the LBO defined in Appendix B. The heat kernel $h_t(x, y)$ relates the amount of heat transferred from point x to point y on the surface after time t. Given the equation:

$$h_t(x, y) = \sum_{l \geq 1} e^{\lambda_l t} \phi_l(x) \phi_l(y) \tag{4.2}$$

The heat kernel fully characterizes shapes up to an isometry and represents increasingly global properties of the shape with increasing time. The heat kernel is invariant under isometric transformations and stable under small perturbations to the isometry.

Since $h_t(x, y)$ is defined for a pair of points over a temporal domain, using heat kernels directly as features would lead to high complexity. Sun et al. [276] proposed using the diagonal of the heat kernel as a local descriptor, referred to as the Heat Kernel Signatures (HKS). HKS restricts itself to just the temporal domain by considering only $h_t(x, x)$. HKS inherits most of the properties of heat kernels under certain conditions. For each point x on the shape, its heat kernel signature is an n-dimensional descriptor (vector) of the form $p(x)$:

$$p(x) = c(x)(K_{t_1}(x, x), \ldots, K_{t_n}(x, x)) \tag{4.3}$$

where $c(x)$ is chosen in such a way that $||p(x)||_2 = 1$. The HKS descriptor has many advantages. First, the heat kernel is **intrinsic** (i.e., expressible solely regarding the Riemannian structure of X), and thus invariant under isometric deformations of X. This makes HKS deformation-invariant. Second, such a descriptor captures information about the neighborhood of a point x on the shape at a scale defined by t. It captures differential information in a small neighborhood of x for small t, and global information about the shape for large values of t. Thus, the n-dimensional feature descriptor vector $p(x)$ can be seen as analogous to the multi-scale feature descriptors used in the computer vision community. Third, for small scales t, the HKS descriptor takes into account local information, which makes topological noise the only local effect. Fourth, Sun et al. [276] prove that, if the LBO of a shape is nondegenerate (i.e., does not contain repeated eigenvalues), then any continuous map that preserves the HKS at every point must be an isometry. This latter property led Sun et al. to call the HKS provably informative.

The computation of the HKS descriptor relies on the calculation of the first eigenfunctions and eigenvalues of the LBO, which can be done efficiently and across different shape representations. Thus makes HKS applicable to different geometric data, as well as triangular meshes.

An extension of this work is a new *intrinsic* **spectral shape descriptors** that are dense and **isometry-invariant** by construction [276].

ShapeGoogle, Bronstein A.M., et al. [38]: A well known feature-based approach in image retrieval is to represent an image as a collection of primitive elements (visual "words") and use the methods from text search such as the "bag of words" paradigm. Then, each image is compactly encoded into a vector of frequencies of occurrences of visual words; a representation referred to as a "bag of features" (BOF) [133]. Images containing similar visual information tend to have same bags of features, and thus comparing bags of features allows retrieval similar

images.

An initial work based on this idea was published initially by Ovsjanikov et al. [211], and later expanded by Bronstein et a.l[38]. The former method uses a feature detector and descriptor based on the heat kernels of the LBO, inspired by Sun et al. [276]. The descriptors are used to construct a vocabulary of geometric words. That is a representation of the shape. This representation is invariant to isometric deformations, robust under a wide class of perturbations, and allows one to compare shapes undergoing different deformations. Traditional quantization and pooling methods were used to generate the bag of features, and a final SVM classifier for classification.

Although the method produced good results, Behmo et al. [19] showed that one of the disadvantages of the bag of features approaches is that they lose information about the spatial location of features in the image, and proposed the commute graph representation, which partially preserves the spatial information. In [38] the authors improved the method adopting an iterative approach based on **dictionary learning**, technique widely used in the computer vision [136]. The method was complemented with a compact representation using binary code indexing and matching with the Hamming distance.

Scale-Invariant Heat Kernel Signatures (SI-HKS), Kokkinos et al [40]: A disadvantage of the HKS is its dependence on the global scale of the shape. If X is globally scaled by β, (i.e. $x' = x//beta$) the corresponding HKS for x' is $\beta^{-2}K_{\beta^{-2}t}(x,x) = \beta^{-2}h_{\beta^{-2}t}(x,x)$.

It is possible in theory to perform global normalization of the shape (e.g., normalizing the area or Laplace-Beltrami eigenvalues), but such a normalization is impossible if the shape has, for example, missing parts. As an alternative, a local normalization was proposed in Bronstein and Kokkinos [40] based on the properties of the Fourier transform. By using a logarithmic scale-space $t = \alpha^{\tau}$, global scaling results in HKS amplitude scaling by β^{-2}, and shift by $2\log_{\alpha}\beta$ in

the scale-space. This effect of scaling is undone by the following sequence of transformations:

$$\begin{aligned} p_{dif}(x) &= (\log K_{\alpha^{\tau_2}}(x,x) - \log K_{\alpha^{\tau_1}}(x,x), \ldots, \\ &\quad \log K_{\alpha^{\tau_m}}(x,x) - \log K_{\alpha^{\tau_{m-1}}}(x,x)), \\ \hat{p}(x) &= |(\mathcal{F}p_{dif}(x))(\omega_1, \ldots, \omega_n)||, \end{aligned}$$

where $\mathcal{F}$ is the discrete Fourier transform, and $(\omega_1, dots, \omega_n)$ denotes a set of frequencies at which the transformed vector is sampled. Taking differences of logarithms removes the scaling constant, and the Fourier transform converts the scale-space shift into a complex phase, which is removed by taking the absolute value. Typically, a large m is used to make the representation insensitive to large scaling factors and edge effects.

Wave Kernel Signature (WKS), Aubry et al. [11]: As the HKS derive from the heat equation model, the Wave Kernel Signature arises from the quantum particle model from the Schrödingers Equation:

$$\left(i\Delta + \frac{\partial}{\partial t}\right)\Psi(x;t) = 0 \tag{4.4}$$

Although similar to the heat equation the induced dynamics are quite different (oscillations rather than mere dissipation).

The main idea of this descriptor is to simulate the behavior of a quantum particle on the manifold possessing some initial energy distribution. The wave function of the particle is given by:

$$\Psi_E(x,t) = \sum_{k0}^{\infty} e^{iE_k t}\phi_k(x) f_E(E_k) \tag{4.5}$$

At time t = 0 the measurement of its energy is E, obtaining an energy probability distribution f_E^2 with expectation E. The probability to measure the particle at a point $x \in X$ is then $|\Psi_E(x,t)|^2$.

The WKS is defined as the average probability (over time) to find a particle in x:

$$WKS(E,x) = \sum_{k=0}^{\infty} \phi_k(x)^2 f_E(E_k)^2 \tag{4.6}$$

since the functions $e^{iE_k t}$ are orthogonal for the L^2 norm. For a better understanding of all the properties of the WKS, we remand to [11]. Here we remark the differences with the HKS. The HKS decomposition in eqn. 4.2 make the HKS composed of low-pass filters, thus is mainly affected by the global shape of the manifold. The WKS, instead, is composed by band-pass filters. This difference is fundamental when we analyze local deformations of the shape.

4.2.1 Generic 3D Shape Retrieval techniques

Shape retrieval [167] is the field where geometry processing techniques have constantly been developed. In this section, we briefly revisit some of the difficulties as well the methods involved. Topology-based methods compare 3D models based on the difference in their global topological structures. Among the various topology representations, Reeb graphs [125], which are rooted in the Morse theory, are considered one of the most popular. View-based techniques use a set of rendered views to represent a 3D model. The visual similarity between the views of two models is regarded as the model difference. A special survey has been published in [178]. Efforts along this line are mostly devoted to two stages: extraction of descriptive features from specific view images, and appropriate comparison between sets of visual features. Recently, Ding and Liu [81] defined a view-based shape descriptor named **Sphere Image** that integrates the spatial information of a collection of viewpoints and their corresponding view features that are matched based on a probabilistic graphics model.

Shape Retrieval. Common tasks in shape retrieval include: intrinsic shape descriptors, shape retrieval, and shape correspondence. Generating intrinsic shape descriptors is the task of producing intrinsic pose and subject-invariant descriptors for human shapes [176]. Ideal descriptors need to have good localization capabilities and discriminative, as well as robust to different kinds of noise, including isometric and non-isometric deformations, geometric and topological noise, different sampling, and missing parts. Shape retrieval [223] is the task to retrieve an object using a query shape, typically after some pose or scale transformation. The job is a hard, fine-grained classification problem since some of the human subjects look nearly identical. Shape retrieval is an established research area with many approaches and methods. For a recent detailed review, see Tangelder and Veltkamp [279].

In Rigid shape retrieval the shape of the object is not subject to any deformation or articulation. Well known methods in this area are: global descriptors based on volume and area [57], wavelets [212], statistical moments [98], self-similarity (symmetry) [145], and distance distributions [210]. Methods reducing the 3D shape retrieval to image retrieval use 2D views [144]; [60]. Graph-based methods based on skeletons [278]; Biasotti et al. [27] are capable of dealing with deformations, for example, matching articulated shapes. Lipman and Funkhouser [175] proposed the Mobius voting scheme for sparse shape matching.

Non-rigid Methods. Unlike generic 3D model retrieval for rigid models, non-rigid 3D model retrieval techniques are dedicated to retrieving the specific and ubiquitous non-rigid 3D models with various poses or articulations. Due to the non-rigid properties of the models, it is more challenging to perform the retrieval. Some important surveys on the topic are [166], and [168]. Despite the elegance and popularity of these spectral methods, they require the input of 3D models to have a manifold data structure, which is unrealistic for most models collected from the web. Therefore, extra preprocessing is needed to remesh the surfaces before feeding them

into the framework.

4.3 Human Body Shape: A Spectral Geometry Approach

In this section, we define the problem and our goal. A human body shape is naturally a non-rigid object that can assume a variety of poses. One property that each shape analysis method "must" have is the pose-invariance. Pose variation is a kind of transformation applied to the mesh. This transformation, which cannot modify the metric of the surface (inelastic deformations of the surface), but only deform it, has been considered an isometry, or in a broad sense a quasi-isometric transformation [237]. As discussed, many of the spectral geometry methods are based on the LBO. Thus these methods are isometry invariant, and for the skinned mesh models pose invariant (see Appendix B.1.1 for the properties of the spectrum). This fact constitutes our fundamental assumption on using spectral geometry techniques for shape analysis problems in soft biometrics, and medical science.

With this assumption, we'll be able to analyze the human body invariant to pose and orientation in a large sense, since we can still use the same spectral content for the shape analysis.

Another benefit of the *intrinsic* characteristic of **spectral analysis** is the invariance of the body parametrization, a fundamental property since we are working on a discretized surface (triangular mesh).

4.3.1 Challenges in non-rigid shape analysis and Spectral Analysis.

Despite the many good properties and the fundamental constraint, working with **spectral geometry** methods have some difficulties. Methods and techniques for spectral geometry have been developed mainly for a triangular mesh. In fact, for this kind of surface discretization, there are some simple and efficient solutions for the LBO operator [276]. Another possible discretization

is the point-cloud. Liu et al. [180] proposed a method for computing the LBO on point-clouds. However, for this kind of discretization K-d tree [24] has to be used to efficiently compute the nearest points [59], before the surface area.

From a feature-based viewpoint work with 3D features is entirely different from the traditional 2D world. The type of invariance in non-rigid shapes is different from one required in RGB images. Typically, feature detectors and descriptors in images are made invariant to **affine transformations**, which accounts for different possible views of an object captured in the image. In the case of nonrigid shapes, the richness of the transformations is much larger, including changes in pose, bending, and connectivity. Since many natural shape deformations (such as articulated motion) can be approximated by isometries, basing the shape descriptors on intrinsic properties of the shape will make it invariant to such deformations. However, shapes are typically less rich in features than images, making it harder to detect a large number of stable and repeatable feature points. This fact poses a challenging trade-off in feature detection between the number of features required to describe a shape on the one hand and the number of features that are repeatable on the other. This dilemma motivates our decision to avoid feature detection all together and use dense descriptors instead.

Unlike images which in the vast majority of applications appear as matrices of pixels, shapes may often be represented as triangular meshes, point clouds, voxels, level sets, etc. Therefore, it is desirable to have local features computable across multiple representations. Finally, since shapes usually do not have a global system of coordinates, the construction of spatial relations between features is a challenging problem.

4.4 WBSA and the Spectrum

In mathematics, especially spectral theory, Weyl's law [8] describes the asymptotic behavior of eigenvalues of the Laplace-Beltrami operator. This behavior is of particular importance since it relates the spectral content of a surface to the surface area.

4.4.1 Weyl's Law on the asymptotic behavior of the eigenvalues.

Let D be a bounded region in $\mathbb{R}^d$, with piecewise smooth boundary B. Let $0 \leq \lambda_1 \leq \lambda_2 \ldots \lambda_n$ be the spectrum, and $N(\lambda)$ the number of eigenvalues $\leq \lambda$, counted with multiplicity. Then

$$N(\lambda) = \frac{vol(D)}{(4\pi)^{d/2}\Gamma(\frac{d}{2}+1)}\lambda^d \qquad \lambda \to \infty \tag{4.7}$$

where $vol(D)$ is the volume of D. For the two common cases we have:

$$\lambda_n \sim \frac{4\pi}{vol(D)}n \text{ for } d = 2 \tag{4.8}$$

and

$$\lambda_n \sim \left(\frac{6\pi^2}{vol(D)}\right)^{2/3} n^{2/3} \text{ for } d = 3 \tag{4.9}$$

Remark: On a surface ($dim(M) = 2$), the Riemannian volume of M is the surface area (A) and the Riemannian volume of the boundary is its length. Then the equation can be written as:

$$\lambda_n \sim \frac{4\pi n}{A} \quad n \to \infty \tag{4.10}$$

or, alternatively:

$$\lim_{n\to\infty} \frac{\lambda_n}{n} = \frac{4\pi}{A} \tag{4.11}$$

In section 4.4.4 we prove the theorem in the case of a rectangular interval. The proof of more general cases is not always easy since we need to evaluate the Dirichlet boundary condition. Before we introduce the new theoretical results let's explain this important theorem in practice.

One of the properties of the LBO spectrum is that is a diverging sequence. What Weyl' formula is saying is that: in general, the eigenvalues asymptotically tend to a line with a slope dependent on the surface area of the 2D manifold. Therefore, a change in the surface area corresponds to a change in the slopes of the eigenvalues asymptotes. Figure 4.1 shows two family of shapes from Virtual NHANES dataset 2.4.1, both females. We can see that for each family the slope change quite a bit; it gets asymptotically very dissimilar. The same behavior is observable for other families of subjects. This makes quite a challenge to compare two shapes. A solution, proposed by Reuter [241] is to normalize the eigenvalues by the surface area to align the spectra.

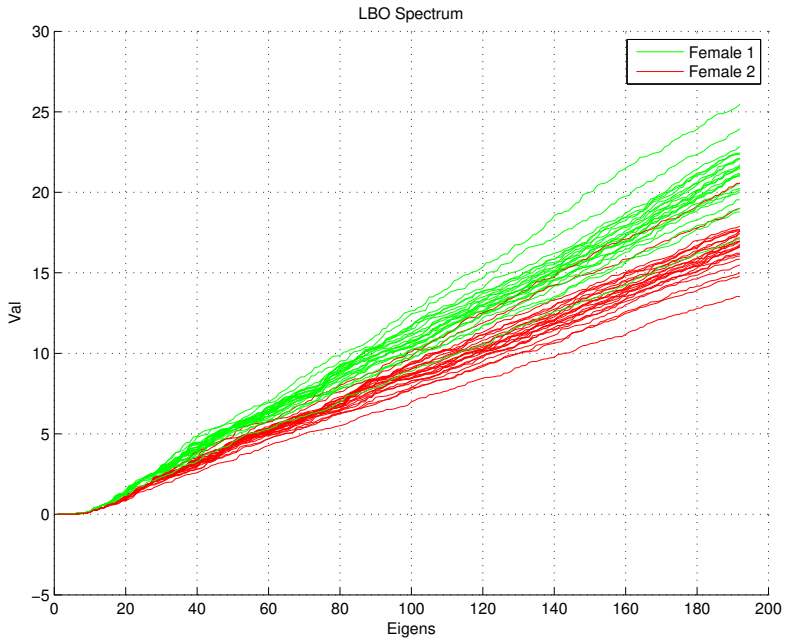


Figure 4.1: LBO Spectrum for two shapes family, females

A quite interesting finding is that the surface area is contained in the spectrum, as the Weyl's

formula showed. This leads to the conclusion that shapes with the same surface area (isometric) can be comparable but is not enough, since the spectrum contain more information about the shape.

4.4.2 LB Spectra of Subdomains

A very interesting problem is the division of the domain D into a finite number of subdomains $D_1, \ldots, D_n$, with each subdomain composed of smooth surfaces $S_1, \ldots, S_n$ (see Figure 4.2). Every subdivision now is a 2D surface, with a border ∂D_i. On these piecewise smooth boundaries we can apply the Dirichlet and Neumann conditions. Given the Helmotz equation:

$$\Delta\psi_n(x) = \lambda_n\psi_n(x) \tag{4.12}$$

with λ_n eigenvalues and $\psi_n(x)$ the eigenfunctions, the boundaries conditions are:

$$\mu_n(x) = 0 \qquad x \in \partial D_i \quad \text{(Dirichlet)} \tag{4.13}$$

$$\frac{\partial}{\partial n}\mu_n(x) = 0 \qquad x \in \partial D_i \quad \text{(Neumann)} \tag{4.14}$$

From the above conditions, there are the Dirichlet eigenvalues $\lambda_1 \leq \lambda_2 \leq \ldots$, and Neumann eigenvalues $\bar{\lambda_1} \leq \bar{\lambda_2} \leq \ldots$. Every subdomain $D_1, \ldots, D_n$ has its own series of eigenvalues. Combining all Dirichlet eigenvalues of all subdomains $D_1, \ldots, D_n$ into a single increasing sequence $\mu_1 \leq \mu_2 \leq \ldots$, and the respective Neumann eigenvalues into another single sequence $\bar{\mu_1} \leq \bar{\mu_2} \leq \ldots$. By the maxmin principle [252], each of these quantities can be obtained as the maximum over piecewise continuous functions $y_1, \ldots, y_{n-1}$ of the minimum over trial functions ω orthogonal to $y_1, \ldots, y_{n-1}$. The trial functions can be defined in all of D simply by

making them vanish in the other subdomains. They will be continuous but not $\mathcal{C}^2$ in the whole domain D. Thus each of the competing trial functions for μ_n has the extra restriction, compared with the trial functions for λ_n, vanishing on the internal boundaries.

It follows that:

$$\lambda_n \leq \mu_n \qquad \text{for each} \quad n = 1, 2, \ldots \tag{4.15}$$

Then, the trial functions defining $\bar{\lambda_n}$ for the Neumann problem in D are arbitrary $\mathcal{C}^2$ functions. All above prove the following Theorem:

$$\hat{\mu_n} \leq \hat{\lambda_n} \leq \lambda_n \leq \mu_n. \tag{4.16}$$

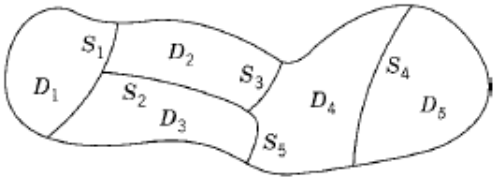


Figure 4.2: Subdomain decomposition.

In a simple 2D case with $D = D_1 \cup D_2 \cup \ldots$, each μ_n corresponds to one of these subdivisions $D_1, D_2, \ldots$. Let $A(D_P)$ be the area of one of the subdivisions, and let $M(\lambda)$ the enumeration function for the sequence $\mu_1, \mu_2, \ldots$ introduced above. Then adding the points which are located within the ellipses, we get:

$$\lim_{\lambda \to \infty} \frac{M(\lambda)}{\lambda} = \sum_p \frac{A(D_P)}{4\pi} \tag{4.17}$$

as for the case of a single rectangle. Since $M(\mu_n) = n$, the reciprocal of 4.17:

$$\lim_{n \to \infty} \frac{\mu_n}{n} = \frac{4\pi}{A(D)} \tag{4.18}$$

and similarly:

$$\lim_{n\to\infty} \frac{\bar{\mu_n}}{n} = \frac{4\pi}{A(D)} \tag{4.19}$$

From theorem 4.16 follow that all the limits are equal: $\lim \lambda_n/n = \lim \bar{\lambda_n}/A(D)$. This result extends theorem 4.11 for the union of 2D domains. The above discussion has been conducted on simplified rectangular subdomains, but it can expanded to more general domains.

4.4.3 Extension to Body Parts

From the above discussion, we obtain some interesting insight on the domain subdivision. This situation can be found in the analysis of body parts. More generally, it is not always possible to get the whole 3D mesh, but only a scanned portion, with a surface Ω ideally smooth, and the respective border $\partial\Omega$. In this situation, it can be possible to extend the spectral analysis to

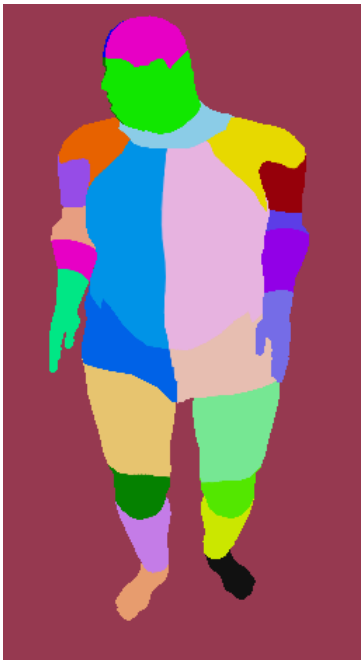


Figure 4.3: Subdomain decomposition in human body parts.

"open" manifolds. Moreover, if we have the body surface subdivided as body parts, the validity

of the Weyl' formula on these subdomains will permit further spectral analyses on the body parts.

Figure 4.3 shows an MH body model with body parts labeled with different colors. This sample has been obtained from the MH engine with a MoCap [67] animation. Extending the WBSA computation to body parts, and the relative Spectral analysis will make the harmonic analysis applicable to situations where a noncomplete mesh is available.

The behavior of the eigenvalues for this setting (partial matching) is highly dependent on the missing portion. LBO eigenvalues are global features of the mesh, and thus not directly suitable for partial matching. Local descriptors (e.g. HKS [276], SIHKS [40]) are better at exploiting the local features. However, if we normalize the eigenvalues with the surface area, as seen in [241], the LBO eigenvalues can still be used for partial shape matching.

4.4.4 Weyl proof for the 2D rectangular interval case

Let us consider the domain $D = \{0 < x < a, 0 < y < b\}$ in the plane. The eigenvalues are of the form:

$$\lambda_n = \frac{l^2\pi^2}{a^2} + \frac{m^2\pi^2}{b^2} \tag{4.20}$$

with the eigenfunctions $\sin(l\pi x/a) \cdot \sin(m\pi y/b)$. Let's introduce the *enumeration function*: $N(\lambda)$ as the number of eigenvalues that do not exceed λ. If the eigenvalues are written in increasing order then $N(\lambda_n) = n$. $N(\lambda)$ can be expressed using 4.20. $N(\lambda)$ is the number of points (l, m) that are contained within the quarter-ellipse:

$$\frac{l^2}{a^2} + \frac{m^2}{b^2} \leq \frac{\lambda}{\pi^2} \qquad (l > 0, m > 0) \tag{4.21}$$

in the (l, m) plane. Each such point is the upper-right corner of a square lying within the quarter ellipse. Therefore, $N(\lambda)$ is at most the area of this quarter ellipse:

$$N(\lambda) = \frac{\lambda ab}{4\pi} \tag{4.22}$$

For large λ, $N(\lambda)$ and this area may differ by approximately the length of the perimeter, which is of the order $\sqrt{\lambda}$. Precisely,

$$\frac{\lambda ab}{4\pi} - C\sqrt{\lambda} \leq N(\lambda) \leq \frac{\lambda ab}{4\pi} \tag{4.23}$$

for some constant C. Substituting $\lambda = \lambda_n$ and $N(\lambda) = n$, we obtain:

$$\frac{\lambda_n ab}{4\pi} - C\sqrt{\lambda_n} \leq n \leq \frac{\lambda_n ab}{4\pi} \tag{4.24}$$

where the constant C does not depend on n. Therefore, dividing by n:

$$\lim_{n\to\infty} \frac{\lambda_n}{n} = \frac{4\pi}{ab} \tag{4.25}$$

the Weyl's law for a rectangle.

4.5 Body Fat Percentage using Spectral Analysis

In this section, we introduce a spectral analysis-based method for estimating the Body Fat Percentage (BFP). As will see ahead, these methods are mostly task driven, where a handcrafted descriptor has to be designed for the specific operation. Taking advantage of the generated VirtualBody dataset in Chapter 2, and the available labels, we develop a spectral method that can classify a given body shape by its BFP by analyzing the harmonic content of the shape. Our

approach benefits from the isometry and intrinsic LBO properties. Moreover, using the scale-invariant Heat Kernel Signature (SI-HKS) [40], the system will be utterly invariant to the scale of the subject. This is one of the fundamental characteristics. Since the body fat percentage is relative to the weight, and not an absolute measure, a non-invariance to scale will make the system biased toward tall subjects.

4.5.1 Problem Definition

In medical science, a common task is the acquisition of some basic measurements, like weight, stature, pressure, etc. For nutritionists, more specific measurements are needed to assess the percentage of body fat. In Chapter 1 we reviewed some of the body composition indicators and the downsides in using the BMI as body fat measure.

Measuring the BFP is a difficult task. Common methods used by physicians are hand measures of the waist and height as recommended by the WHO [130], fat calipers, scale with bio-electrical impedance analysis, and bodpod analysis (Figure 4.4). The bioelectrical impedance analysis is the only cost-effective automatic method. It permits to have a measure of the BFP, BMI, weight, and water in the body in a matter of seconds, just stepping on the scale, everyone at home can use it. However, the accuracy is sometimes not good. This method is highly affected by the water in the body, and the skin conductivity.

Hand measurements, calipers, and bodpod provide accurate measurements, but the methods need trained physicians or a lab technician for the bodpod. The bodpod is also an expensive machine (Figure 4.4). The recurring use of a physician is often costly and not feasible when the subjects live in remote areas. Self-assessment can often be biased by the individual, and cannot be used as a reliable measure. Moreover, monitoring the body composition over time is an important task and needs to be made a lot easier and cost-effective. A reliable, cost-effective, automatic system will make the prevention and monitoring of obesity much easier.

Figure 4.4: BodPod setup. Courtesy of lorainccc.edu.

A system with the above specification can be used without major constraints as a soft biometric. Weight prediction has been proposed as soft biometric feature in [49],[3],[290],[291]. Similarly, BFP can be used as a soft biometric feature if it can be easily detected.
For the above reasons a system capable of a fast and reliable estimate of the BFP is valuable for many disciplines and research areas.

We present a system that can take advantage of modern 3D acquisition systems and techniques in **spectral analysis** to estimate the BFP for humans. This system will benefit from the pose-invariant nature of the spectral techniques, making the approach immune from the usual anthropometric measurement problems (Chapter 1).

4.5.2 Proposed method

Given a 3D acquisition of the body subject, our goal is to detect in which BFP class he/she falls. The World Health Organization (WHO) defines some predictive values regarding Waist-to-Height ratio (WHR) [9] and BMI as related to cardiovascular and weight-related diseases [262]. In particular, the WHO recommended being extremely careful when the WSR is greater to 0.6. Table 4.1 shows values of WHR and their corresponding classification [9].

Using the VirtualBody dataset presented in Chapter 2 we develop a system to categorize

Children (< 15)	Men	Women	Categorization
< 0.34	< 0.34	< 0.34	Extremely Slim
0.35 to 0.45	0.35 to 0.42	0.35 to 0.41	Healthy Slim
0.46 to 0.51	0.43 to 0.52	0.42 to 0.48	Healthy
0.52 to -.63	0.53 to 0.57	0.49 to 0.53	Overweight
0.64 +	0.58 to 0.62	0.54 to 0.57	Very Overweight
	0.63+	0.58+	Morbidly Obese

Table 4.1: WHR values and relative categorization [9].

subjects. For simplicity, we define three classes: **lean**, that corresponds to the healthy slim class of Table 4.1, then **average**, and **fat** classes, for respectively, healthy and overweight classes of Table 4.1.

We group the subjects from the VirtualBody dataset in these three classes, by considering their WHR values independent of age. This is a critical design decision. We want to be robust with respect to age. This, in practice, is a scale-invariant problem. Children can be as tall as 120 centimeters in our dataset, while adults, can be more than 2 meters. This is a significant difference considering the dimensionality of each mesh model. To create an invariant system will be more difficult and requires more sophisticated techniques.

4.5.3 Interaction between BFP and Body Weight.

The Body Fat Percentage is a relative measure of the body mass portion constituted by fat. The interaction of BFP and weight on the visual appearance is not well understood. This interaction is unfortunately nonlinear and quite complex. Moreover, different categories: males, females, as well as different races and age groups have significantly different patterns.

The weight of a subject can be considered as a global measure of the shape. When weight increases, the total shape change, although, changes in weight could be due to different factors. From the early age of childhood till the adulthood, the typical growth is the primary cause of

change in weight, thus shape. The shape changes mainly due to the stature, but it can also change due to an increase in fat when healthy habits are not followed. Pediatricians and nutritionists have the famous growth charts to monitor a correct growth. However, waist or hip measures are not uncommon, since these are the areas where an accumulation of fat is most probably to occur. Then, we can conclude that: in general, concerning the visual appearance, subjects with low BFP have a shape that changes globally (mostly in the stature dimension) with weight increases, while those with high BFP exhibit more local changes of their shape (waist, hips, torso).
With this in mind, accurately classify people by their BFP, the system needs to be able to detect these local changes in body shape with changing weight.

To better understand the change of the small BFP variations in the shape, it is indispensable to have a dataset with enough variations and descriptive labels. The Virtual NHANES collection 2.4.1 of the VirtualBody dataset presented in Chapter 2 is designed with this in mind. The Virtual NHANES population (Section 2.4.1) is composed of families of shapes of the same subjects, with the same stature but with variations in weight and fat percentage. These labels and corresponding shapes will be extremely useful in the subsequent learning stage.

Apparently, the proposed method is similar to the well-known shape retrieval task [37]. However, our dataset presents more challenges. As discussed in Chapter 2, Tosca [36], Scape [7], the new FAUST dataset [31], and the SHREC'10 datasets [37] are quite challenging, but the number of subjects and the nuisances are designed to test some properties, e.g., invariance in pose (isometry transformation) and different kind of noise (topological, holes, remeshing, etc). Our newest dataset is completely different since it presents challenges very different than the previous dataset.

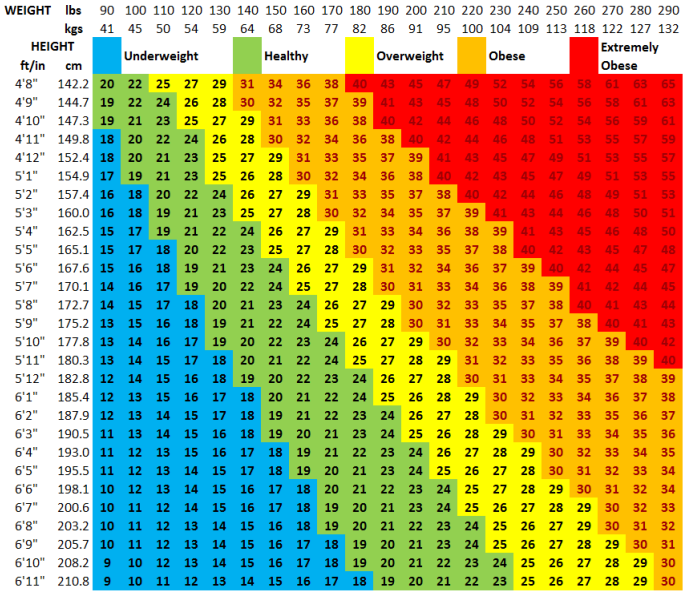
To understand the data we need to introduce some important concepts in human body composition. The two main variables are weight and BFP. Since BFP changes the weight, the effects

on the visual appearance for low weight subjects is very different with respect to high weight. Increasing the weight, the BFP effect on the shape is greater (e.g., a very little effect can be seen on the shape when the subject is anorexic). Moreover, biologically it is unlikely to have a body composed of zero or 100 % of fat. This natural behavior is correctly interpreted in the MakeHuman engine [16], and confirmed in the virtual NHANES dataset 2.4.1.

To further investigate and quantify this behavior, we measure the Hausdorff distance [245] between the meshes. The results are shown as graphs in Figures 4.7, 4.9. Each graph represents a family of meshes for one individual in the Virtual NHANES 2.4.1 dataset. In the dataset, there are 25 variations of the same subject (in the graph only some are shown). The variations are obtained by changing weight and BFP. The graphs can be considered as having an origin at the bottom left corner. The Y-coordinate corresponds to the weight dimension, while the X-coordinate corresponds to the BFP dimension. Each node corresponds to one of the 25 variations on a subject shape. The value at each edge is the Hausdorff distance between two shapes.

Shapes of the same weight, are along the X-coordinate. Shapes with the same BFP are aligned along the Y-coordinate. From this representation we can verify the above assertions. If we consider shapes with maximum weight, (top row W1), the Hausdorff distance between the state with low BFP and high BFP is quite high: subject 10 Male (Figure 4.7) at W1 $dist(M0, M1)_{w1} = 1,278$. Let's consider shapes with minimum weight, bottom row W0. The distance between shapes with the same BFPs: $dist(M0, M1)_{w0} = 0.460$. Thus $dist(M0, M1)_{w1} > dist(M0, M1)_{w0}$, for the same BFP values: M0,M1.

A second very interesting effect is the behavior of the **average** and **lean** subjects. These shapes fall on the right portion of the graph. If we consider these shapes we can see that the Hausdorff distance is quite small. This corresponds with the natural behavior, subjects in good health with right BFP are more difficult to detect since they look alike. These situations can be



WEIGHT	lbs	90	100	110	120	130	140	150	160	170	180	190	200	210	220	230	240	250	260	270	280	290
	kgs	41	45	50	54	59	64	68	73	77	82	86	91	95	100	104	109	113	118	122	127	132
HEIGHT		Underweight					Healthy				Overweight				Obese				Extremely Obese			
ft/in	cm																					
4'8"	142.2	20	22	25	27	29	31	34	36	38	40	43	45	47	49	52	54	56	58	61	63	65
4'9"	144.7	19	22	24	26	28	30	32	35	37	39	41	43	45	48	50	52	54	56	58	61	63
4'10"	147.3	19	21	23	25	27	29	31	33	36	38	40	42	44	46	48	50	52	54	56	59	61
4'11"	149.8	18	20	22	24	26	28	30	32	34	36	38	40	42	44	46	48	51	53	55	57	59
4'12"	152.4	18	20	21	23	25	27	29	31	33	35	37	39	41	43	45	47	49	51	53	55	57
5'1"	154.9	17	19	21	23	25	26	28	30	32	34	36	38	40	42	43	45	47	49	51	53	55
5'2"	157.4	16	18	20	22	24	26	27	29	31	33	35	37	38	40	42	44	46	48	49	51	53
5'3"	160.0	16	18	19	21	23	25	27	28	30	32	34	35	37	39	41	43	44	46	48	50	51
5'4"	162.5	15	17	19	21	22	24	26	27	29	31	33	34	36	38	39	41	43	45	46	48	50
5'5"	165.1	15	17	18	20	22	23	25	27	28	30	32	33	35	37	38	40	42	43	45	47	48
5'6"	167.6	15	16	18	19	21	23	24	26	27	29	31	32	34	36	37	39	40	42	44	45	47
5'7"	170.1	14	16	17	19	20	22	24	25	27	28	30	31	33	34	36	38	39	41	42	44	45
5'8"	172.7	14	15	17	18	20	21	23	24	26	27	29	30	32	33	35	37	38	40	41	43	44
5'9"	175.2	13	15	16	18	19	21	22	24	25	27	28	30	31	33	34	35	37	38	40	41	43
5'10"	177.8	13	14	16	17	19	20	22	23	24	26	27	29	30	32	33	34	36	37	39	40	42
5'11"	180.3	13	14	15	17	18	20	21	22	24	25	27	28	29	31	32	33	35	36	38	39	40
5'12"	182.8	12	14	15	16	18	19	20	22	23	24	26	27	28	30	31	33	34	35	37	38	39
6'1"	185.4	12	13	15	16	17	18	20	21	22	24	25	26	28	29	30	32	33	34	36	37	38
6'2"	187.9	12	13	14	15	17	18	19	21	22	23	24	26	27	28	30	31	32	33	35	36	37
6'3"	190.5	11	13	14	15	16	18	19	20	21	23	24	25	26	28	29	30	31	33	34	35	36
6'4"	193.0	11	12	13	15	16	17	18	19	21	22	23	24	26	27	28	29	30	32	33	34	35
6'5"	195.5	11	12	13	14	15	17	18	19	20	21	23	24	25	26	27	28	30	31	32	33	34
6'6"	198.1	10	12	13	14	15	16	17	18	20	21	22	23	24	25	27	28	29	30	31	32	34
6'7"	200.6	10	11	12	14	15	16	17	18	19	20	21	23	24	25	26	27	28	29	30	32	33
6'8"	203.2	10	11	12	13	14	15	16	18	19	20	21	22	23	24	25	26	27	29	30	31	32
6'9"	205.7	10	11	12	13	14	15	16	17	18	19	20	21	23	24	25	26	27	28	29	30	31
6'10"	208.2	9	10	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
6'11"	210.8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	25	26	27	28	29	30

Figure 4.5: BMI chart.

quite challenging to overcome. This challenge is at the center of a recent topic called **metric learning**.

Although the exciting graph representation (Figures 4.7 to 4.9), where each node has the same probability to occur, it does not consider the probability that a certain shape can occur. In fact, the generation of the subjects in Chapter 2 has been done independently from the statistics of the population but just using some measurements and the MH morphing capabilities. This assumption is totally fine for designing and testing a new approach because the algorithm needs to be robust to a greater number of transformations. However, a correct representation of the families of shapes is to account for the probability that a certain state (shape) in the diagram can occur. Useful prior information can be extracted from an accurate statistical analysis of the NHANES dataset [56]. Other useful information can be the BMI charts in use by physicians (Figure 4.5) A statistical model often applied in pattern recognition and machine learning, used for structured prediction is the Conditional Random Fields (CRFs)[155]. CRFs are a type of discriminative undirected probabilistic graphical model. It is used to encode known relation-

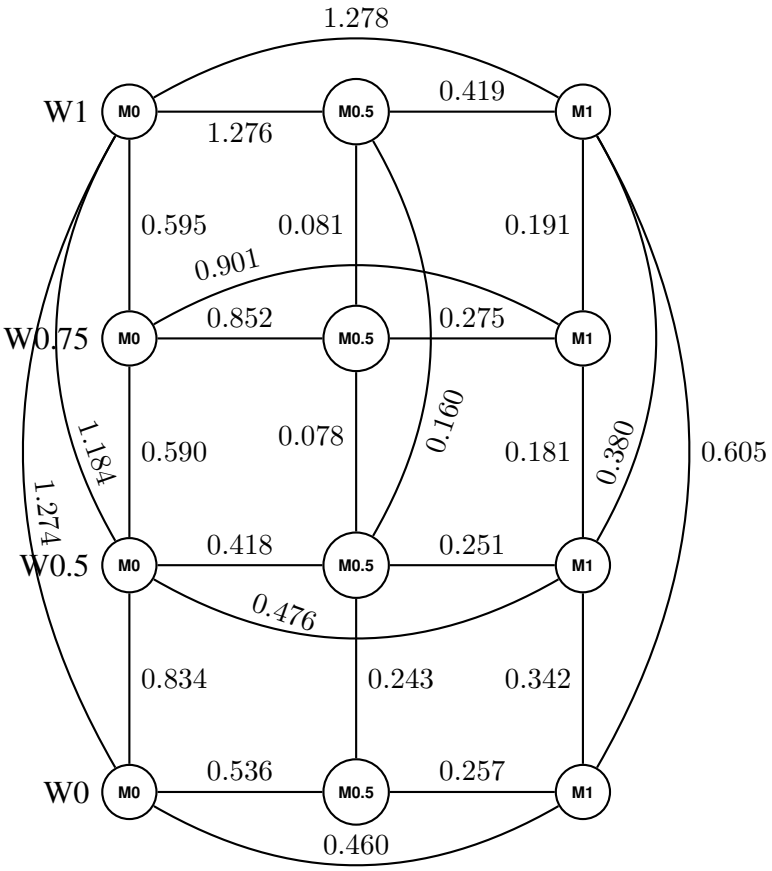


Figure 4.6: Interaction between body weight (y-axis) and BFP (x-axis). Each node represents the body shape generated by varying the weight (W0-W1), and BFP (M0-M1) of the average subject located at (W0.5,M0.5). Edges represent the Maximum Hausdorff distance between the body shapes at the associated nodes. Results shown only 11 variations of Subject 10, a male subject.

ships between observations and construct consistent interpretations. It is often used for labeling or parsing of sequential data, such as natural language text or biological sequences [155] and in computer vision [119][62].

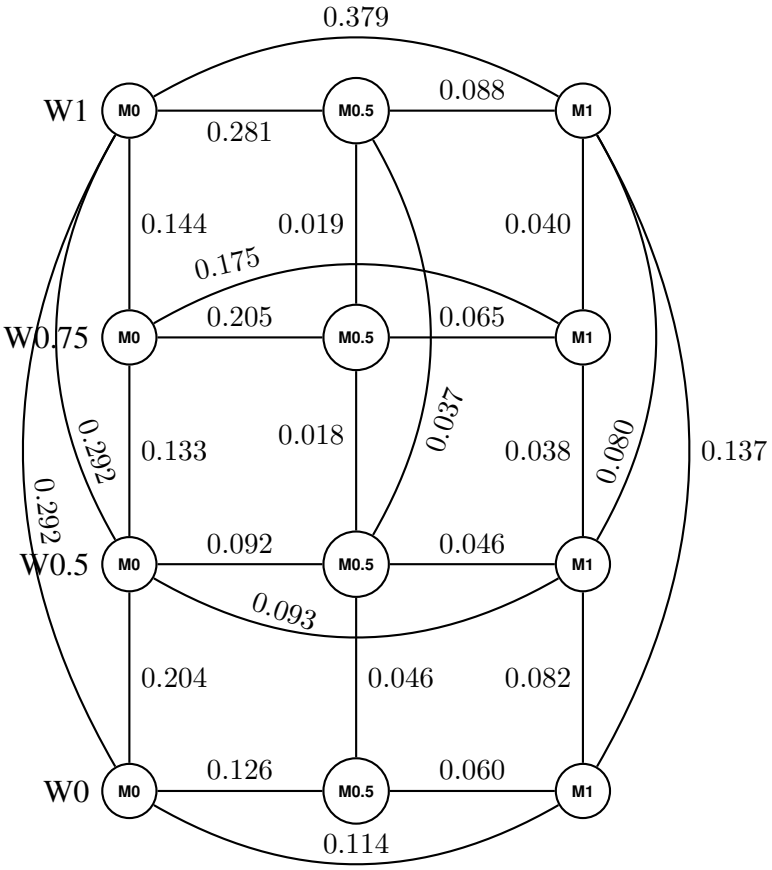


Figure 4.7: Interaction between body weight (y-axis) and BFP (x-axis). Each node represents the body shape generated by varying the weight (W0-W1), and BFP (M0-M1) of the average subject located at (W0.5,M0.5). Edges represent the RMS Hausdorff distance between the body shapes at the associated nodes. Results shown only 11 variations of Subject 10, a male subject.

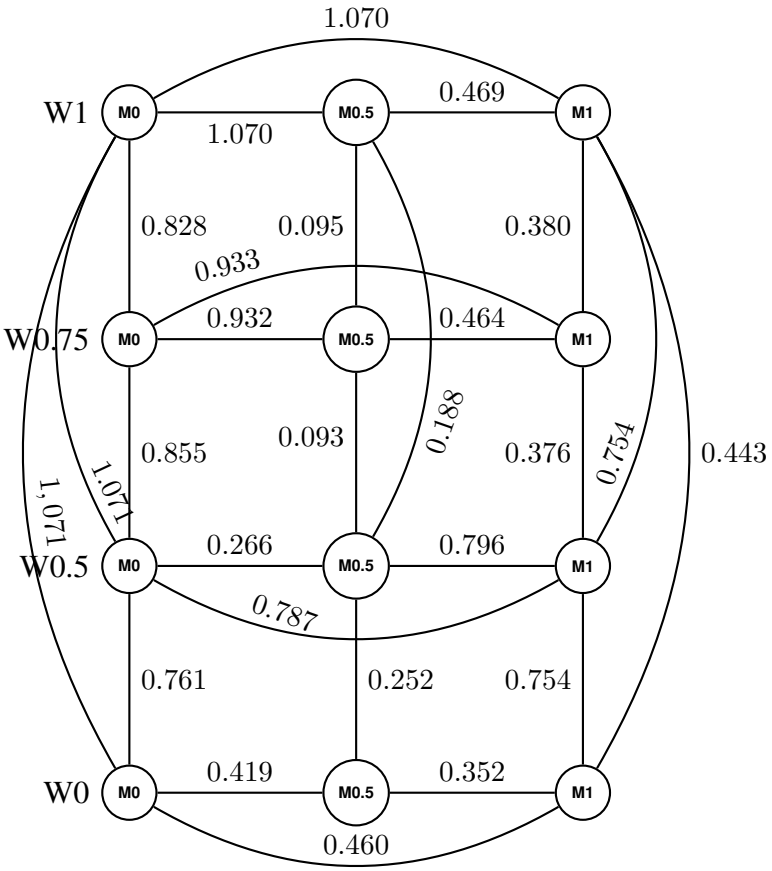


Figure 4.8: Interaction between body weight (y-axis) and BFP (x-axis). Each node represents the body shape generated by varying the weight (W0-W1), and BFP (M0-M1) of the average subject located at (W0.5,M0.5). Edges represent the Maximum Hausdorff distance between the body shapes at the associated nodes. Results shown only 11 variations of Subject 7, a female subject.

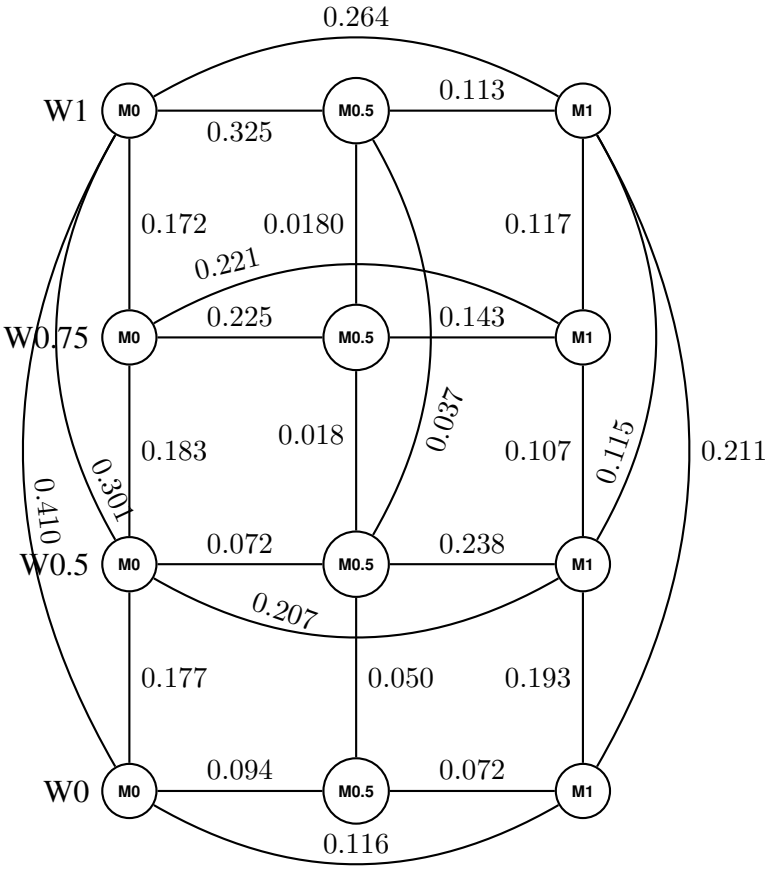


Figure 4.9: Interaction between body weight (y-axis) and BFP (x-axis). Each node represents the body shape generated by varying the weight (W0-W1), and BFP (M0-M1) of the average subject located at (W0.5,M0.5). Edges represent the RMS Hausdorff distance between the body shapes at the associated nodes. Results shown only 11 variations of Subject 7, a female subject.

4.5.4 Bag of Features Approach

Given the triangular mesh, modeled as a two-dimensional manifold S, and sampled at n points $s_1, \ldots, s_n$. The next step is to compute the local descriptors $\mathbf{x}(\mathbf{s_i})$. We denote the LBO of S as ΔS, and we use the cotangent method [226] to obtain the discretized version. This discretization preserves many important properties of the continuous Laplace-Beltrami operator, such as positive semidefiniteness, symmetry, and locality. The eigenfunctions and eigenvalues of the LBO $\Delta S \phi_l = \lambda_l \phi_l$ are denoted $\{\phi_l, \lambda_l\}_{l \geq 1}$ where, $\lambda_i \quad i = 1, \ldots$ are the eigenvalues, and $\phi_i \quad i = 1, \ldots$ the eigenfunctions.

The features used in this work are obtained from the scale-invariant Heat Kernel Signature [40]. The HKS descriptor [276] $h_t(s_i, s_j)$ has many advantages, but unfortunately is dependent on the global scale. It can be made scale-invariant using the technique discussed in Section 4.2. Such invariance is critical in our case since we deal with meshes at different scales: kids and women are typically smaller, and the invariance concerning the subject weight is related to the invariance of the global shape. Figure 4.10 shows the variance of the LBO eigenvalues for two families of subjects. As discussed before, the variations in the family are due to weight and BFP. Given a fixed stature, the shape deformations are localized mainly where the fat is stored (waist, torso, etc.), and the muscle bulging is more evident. As we can see from Figure 4.10, in these areas the LBO response is very high. Instead, there are some areas where the variation is very low, e.g., face, feet, hands. We believe the MH engine does not model well this body parts against BFP and weight variation. This, however, it does not affect our study, which is focused on the overall body shape, and not on single body parts.

In this work, we adopt the Bag of Features (BOF) method [133], a popular approach in computer vision. Given a set of q-dimensional descriptors at all the n points of the shape, we

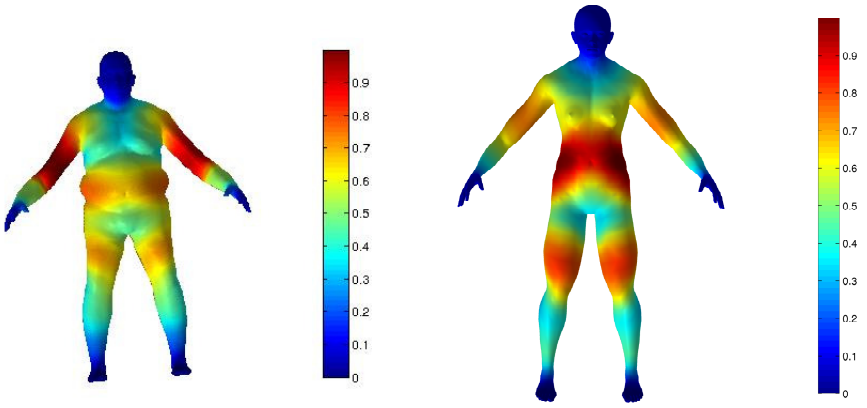


Figure 4.10: Area Variance: (Left) Short Male, (Right) Muscular Female.

represent the shape as $q \times n$ matrix:

$$\mathbf{X} = (\mathbf{x_1}, \ldots, \mathbf{x_n}) = (\mathbf{x}(\mathbf{s_1}), \ldots, \mathbf{x}(\mathbf{s_n})). \tag{4.26}$$

A Bag-Of-Features is a global descriptor composed by quantized elements in a geometric dictionary and then computing the frequency of these geometric words.

A **geometric** dictionary is a $q \times v$ matrix $\mathbf{D} = (\mathbf{d_1}, \ldots, \mathbf{d_v})$, whose columns are descriptors, called "geometric words", or atoms, where v is the dimension of the dictionary.

The dictionary is constructed offline using a large collection of shapes, clustering the respective descriptors (q-dimensional space) into v Voronoi regions, using the k-means algorithm.

Quantization. Given a dictionary $\mathbf{D}$, each local descriptor $\mathbf{x}(\mathbf{s_i})$ is replaced by the closest word in the dictionary:

$$\mathbf{Z}(\mathbf{X}, \mathbf{D}) = arg \min_{i=1,\ldots,v} ||\mathbf{x} - \mathbf{d_i}||_2 \tag{4.27}$$

This process is called *vector quantization*. Integrating the feature distribution over the entire

shape S we obtain a vector $n \times 1$:

$$f(X) = \int_s \mathbf{X}(\mathbf{s}) d\mu(s) \tag{4.28}$$

which is called **Bag of Features**. The operation of integral is intended as "pooling" together with the contributions of the different local features. This operation, executed over the entire shape, makes the representation insensitive to the spatial locations of the features. However, if we randomly change the position of the features, but keeping the same distribution the system will give precisely the erroneous same result. In case of shapes, this phenomenon may be even more pronounced, as shapes, being poorer in features, tend to have many similar geometric words. The analogy of expressions in shapes would be spatially-close geometric words.
Instead of looking at the frequency of individual geometric words, a better approach will be to consider the frequency of word pairs, thus accounting not only for the frequency but also for the spatial relations between features.

Overall, the former representation is very convenient since comparing two shapes is just matter of a simple operation:

$$d_{BOF}(X, Y) = ||f(X) - f(Y)||_1 \tag{4.29}$$

Classification. Our ultimate goal is to classify the shapes with respect to Body Fat Percentage (BFP). This classification task is significantly different from the shape retrieval framework, where spectral distances are computed to retrieve the shape. The BOFs, together with the labels are used to train a classifier. We use the SVM classifier [33]. SVMs have become the method of choice to solve difficult classification problems in a wide range of application domains [17][287][124][280]. Training involves optimization of a convex cost function, and there are no false local minima to complicate the learning process. SVM has many benefits, for

example, the model constructed has an explicit dependence on the most informative patterns in the data (the support vectors). Hence interpretation is straightforward. SVMs are a well-known class of algorithms which use the idea of kernel substitution, making them able to find non-linear solutions.

In the general case of two class problem, given a set of instance-label pairs $(\boldsymbol{x_i}, y_i), i = 1, \ldots, l \quad \mathbf{x} \in \mathbb{R}^n, y_i \in \{+1, -1\}$ the method solve the following unconstrained optimization problem with loss functions $\xi(\boldsymbol{\omega}; \boldsymbol{x_i}, y_i)$:

$$\min_{\boldsymbol{\omega}} \frac{1}{2}\boldsymbol{\omega}^T\boldsymbol{\omega} + C\sum_{i=1}^{l}\xi(\boldsymbol{\omega}; \boldsymbol{x_i}, y_i), \tag{4.30}$$

where $C > 0$ is a penalty parameter. For SVM, the two common loss functions are $\max(1 - y_i\boldsymbol{\omega}^T\boldsymbol{x_i}, 0)$ and $\max(1 - y_i\boldsymbol{\omega}^T\boldsymbol{x_i}, 0)^2$. The former is referred to as L1-SVM, while the latter is L2-SVM. One can show that the solution has the form:

$$\hat{\boldsymbol{\omega}} = \sum_i \alpha_i \boldsymbol{x_i} \tag{4.31}$$

where $\alpha_i = \lambda_i y_i$. The $\boldsymbol{x_i}$ for which $\alpha_i > 0$ are called support vectors; these are points which are either incorrectly classified, or are classified correctly but are on or inside the margin.
The multi-class problem is formalized as a One-Vs-The-Rest strategy. This strategy, also known as **one-vs-all**, consists in fitting one classifier per class. For each classifier, the class is fitted with all the other classes. The samples of that class are the positive samples and all other samples as negatives. This strategy requires the base classifiers to produce a real-valued confidence score for its decision, rather than just a class label; discrete class labels alone can lead to ambiguities, where multiple classes are predicted for a single sample [28].

4.6 Results

To assess our method, we performed shape classification. We used the human body shapes from The Virtual NHANES dataset and subdivided these into three classes: **Lean**, **Fat**, and **Average**, as described before. The basic method: after features extraction using the scale invariant Heat Kernel Signature (siHKS), compute the dictionary using the k-means algorithm. The quantized signature is then used to train an SVM classifier.

4.6.1 Dataset Preparation

Our task is subject categorization based on the BFP, invariant to the subject weight. We partition the dataset as shown in Figure 4.11, where, each category is composed of subjects with different weights.

From the above discussion on shape distances vs. weight, we repeat different experiments with various weight groups. The first experiment is using shapes with weight in the range (W0.5-W1). The second experiment considers (W0-W1). We repeat these experiments for males and females subjects. We also create another significant partition. The HKS and thus the siHKS are local descriptors, thus are insensible by stature variations. However, the BOF framework is positively affected by this variation as we discussed above. With this experiment we expect to find some decrease in performance.

4.6.2 siHKS Features

The scale-invariant Heat Kernel Signature has two main parameters, time intervals, and sampled frequencies. The time intervals determine the spatial frequencies analyzed. The second parameter determines the sampled frequencies from the LBO eigen-decomposition. We use a relatively small number of frequencies (from 2 to 20 with a step of 0.2), considering only the

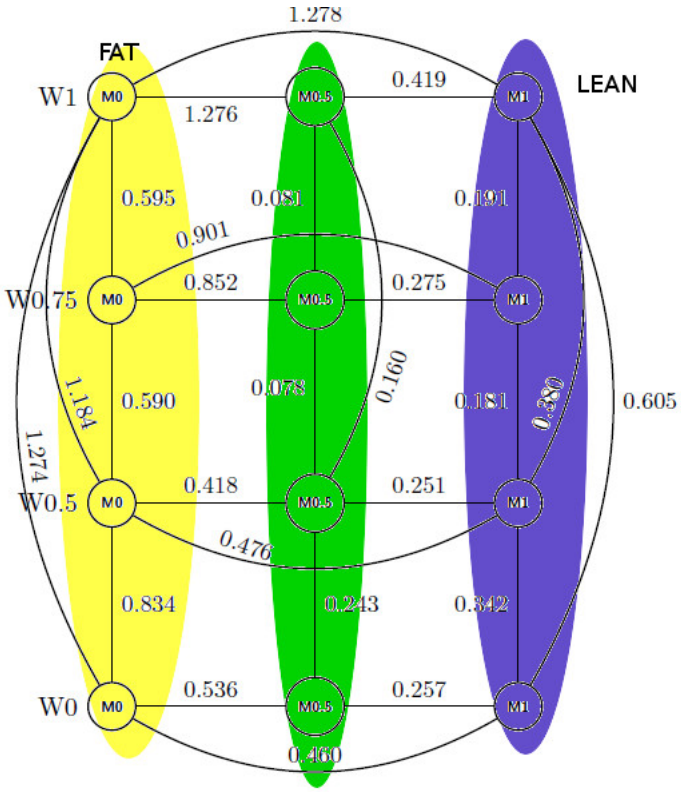


Figure 4.11: Categorization of subjects in VirtualBody dataset. Yellow for Fat, Green for Average, Blue for Lean.

first 20 components from the LBO.

4.6.3 Training

For all the experiments we split the samples in the ratio 70/30 between training and testing. For the experiments with limited weight, we used 240/104 shapes for class, for train and testing respectively. For the experiment with all the weights classes, we used a more substantial number: 700 and 490. We decide to use a linear SVM from the liblinear library [88] to keep the training time reasonably low, but we use the homogeneous kernel map [289] to take advantage of additive kernels, in our case χ^2.

4.6.4 Performance

The results of the classification are shown in Table 4.2 in terms of classification accuracy. As expected, using all the weight variations the system produced a low performance. This behavior was expected since the basic dictionary learned is not capable of accounting for the complex distances between groups. We believe that with a more refined discriminative-generative learning technique we can obtain significant improvements. For instance, using a different pooling strategy, like tf-idf, or in general, learn the "pool" operator. Constraining the shape variations limiting the stature and age (global variations) we can see that the performances increase significantly. From Table 4.2 we notice an unexpected difference in the results for males and females. The phenomenon is because the female shape is more complex and the variations are harder to describe [234]. Unfortunately, we cannot compare this results with any other work since this is the first of this kind. Works on body weight estimation cannot be easily compared with the present without making some drastic unfair assumptions. Table 4.2 reports the classification accuracy for two clusters: all body, and adults. We can observe that for childs and small bodies, the above assumption on females shapes do not hold. In this scenario the two groups (males,females) are very similar, and the experiment is not conclusive since the two classes can be unbalaced. Unfortunately, for these subjects the shape information is not enough descriptive.

Results	Males		Females	
	Weights W0-W1	Weights W0.5-W1	Weights W0-W1	Weights W0.5-W1
All statures, all ages				
Accuracy	71.93 %	77.19 %	84.93 %	86.32 %
Stature $>$ 140 cm. $>$ 15 yrs old				
Accuracy	95.49 %	96.39 %	89.74 %	89.10 %

Table 4.2: Classification results.

4.7 Conclusion and Future Work

In this chapter we have introduced the Spectral Geometry (SG) to the problem of human body description. SG permits to describe non-rigid objects under heavy isometric deformations. This characteristic is particularly interesting for developing a robust description of the body under pose variation.

Taking advantage of the above theory we have presented an original work on the BFP estimation. The present method is based on **spectral geometry** techniques, robust to body pose and deformations, based on the *intrinsic* description of the shape, thus invariant to any reference system and camera orientation. The algorithm has been developed using the new introduced dataset for body shape analysis, permitting the creation of very challenging conditions.

We also exploit some new, and well-known limits using the traditional BOF framework. Learning the dictionary in an unsupervised manner has its limits and is one of the areas where we can make new contributions. A possible solution is to learn a sparse dictionary in a supervised manner. There are many recent works on this topic [12], and the area of dictionary learning with sparse representation is moving quite fast.

We propose to model the transformations of the shape when it is subjected to weight variations, but contained in the same BFP group. Inspired by the work [211] on functional maps, we can learn the deformations the shape can sustain with weight variations. Thus decoupling the variations due to the weight from the changes due to the BFP. Another contribution is to learn the dictionary using spectral analysis techniques. Using local spectral descriptors (HKS, siHKS, WKS, etc.), we can learn the optimal transfer function [176] for the learning process.

A future study to obtain a better representation of families of shapes would be the use of probabilistic graphical models (PGM) or conditional random fields (CRF). In particular, we want to infer some useful prior knowledge in the graph representation, modeling the shape

transformations as state transitions in the CRF model.

Chapter 5

Pose Invariant Soft Biometrics

A soft biometric [169],[135] is defined as any anatomical or behavioral characteristic that provides some information about the identity of a person, but may not be sufficient to identify the subject. Gender, ethnicity, age, height, weight, eye color, scars, marks, tattoos, and voice accents are typical soft biometric traits. In particular, the anatomical characteristics of the human body, in the form of anthropometric measures: height, waist circumference, torso, etc., constitute the geometric description of the body, that we call *anthropometric soft biometrics*. The first biometric system, established by Alphonse Bertillon in 1883 [135], long before the notion of soft biometric, was based on anthropometric measurements, and other soft-biometric traits like tattoo and scars. Subsequently, with the introduction of **hard** biometrics (e.g., fingerprint, iris, face, etc.), the **soft** biometrics has been used mainly to complement the former to improve recognition accuracy [135]. Recently, soft-biometrics has taken a life on his own with the advent of modern surveillance systems, long-range cameras, and the recent consumer products such as Microsoft Kinect [71], and Intel RealSense [197]. These permit to easily acquire traditional natural images, as well the additional geometric information. Due to their compact form factor, these can be used in many settings, previously inaccessible to traditional cumbersome

laser scanners. Moreover, the recent advances in sensing technology applied to self-driving cars, robotics, and drones (e.g., solid state lidar [70]), has ushered new environments where 3D data assumes a vital role in the detection, recognition, and avoidance of objects, and most importantly humans [205]. In these new settings, there is an increasing demand for more powerful and fast algorithms, capable of efficient representation and utilization of the geometric information of the human body. Unfortunately, anthropometric measurements, relying on traditional body part measures are becoming inadequate, and restrictive for these recent applications, where partial bodies in unconstrained pose are acquired from random views. Traditionally, anthropometric measurements have been acquired manually, and the existing collections of anthropometric data, such as CAESAR [244], NHANES [56], and ANSUR [86] are based on elaborate methodologies for hand measuring human bodies. Leveraging computer vision techniques to automatically extract these measures has been investigated before. However, simple automation without introducing novel and reliable descriptors is restrictive and akin prone to creating a computerized version of the manual hand-measuring process, without exploiting the real machine vision capabilities.

A key challenge, often ignored by the previous work on anthropometric measurements is the problem of the pose. The human body can assume a high number of poses (see Fig. 5.1), and can take an equally high number of body shapes, due to the non-rigid nature of its composition (e.g., fat/lean ratio). This has a significant impact on how we extract the measurements, and how we represent them. While the problem of pose has been recognized in other related areas, such as computer vision [80], and body shape modeling [115, 127, 227], it has been largely ignored in anthropometric soft biometrics, which typically assumes the person is constrained to a certain standard pose [3, 22, 188, 244, 291]. In this work, we investigate the pose problem performing a detailed statistical analysis of the anthropometric measurements under pose variation show-

ing their dispersive behavior. With the increasing use of anthropometric soft biometric in more unconstrained scenarios, we must find ways to address the challenge of the body representation under pose variation. We approach the anthropometric body description recurring to spectral geometric techniques, able to represent non-rigid objects under non-Euclidean transformations. The main contribution is the comparison of the new pose invariant representation against the traditional anthropometric measurements under pose variation. This new study has been possible introducing a new dataset of virtual subjects, containing anthropometric measurements, as well as 3D data for each pose assumed by the subjects. This novelty constitutes a huge advantage not only in anthropometric soft biometrics, but also in medicine, robotic vision, and in all the applications which use RGB-D, and lidar [112, 179, 303] devices, where a geometric pose invariant description of the body is needed. Interestingly, the present method can be used in a more general framework as a labeling stage for training new machine learning algorithms, similar to the Kinect body tracker algorithm [266], that made the success of RGB-D devices.

A Common problem in representation learning is the description of semantically meaningful quantities, like the anthropometric measurements, with non-human interpretable descriptors, like the new description. We propose a simple method able to predict traditional anthropometric measures using the new spectral geometric representation, bridging the gap between the two descriptions. The proposed solution differs from the previous works in anthropometric soft biometrics since it makes use of recent innovations in body modeling, spectral geometry, computer graphics, and machine learning. In Section 5.1, we provide a background to the work, and describe major related efforts in soft biometrics, with a focus on 3D body shape. In Section 5.2, we perform a new statistical analysis to show the performance loss of traditional anthropometric measurements under pose transformations. Section 5.3 introduces our proposed spectral geometry approach, to address the problem of pose, and the semantic predictor. In Section 5.4.1

we describe the new dataset method to label the geometric information of human bodies. Section 5.4 presents experimental results using both real and synthetic datasets.

Figure 5.1: The 18 poses in the Virtual Pose Dataset (VPD), (from left to right, top to bottom): Benchmark, Default, Fight1, Standing6, Fight2, Fight3, Fly1, Fly2, Fight4, Standing3, Gym1, Tpose, Standing5, Run1, Standing1, Standing2, Sit1, Standing4.

5.1 Background and Literature Review

Recently, there has been an increased interest in soft biometric features, where the robust extraction of these features is still an open problem. When traditional biometric features are available, soft biometric traits can be extracted more efficiently. For example, given the face image, various attributes can be extracted with sufficient reliability, e.g., gender [50], ethnicity [111], age [154, 156], and eye color. However, the need for the primary biometric features is a key limita-

tion. Soft biometric systems are reviewed in recent surveys by Dantcheva et al. [77, 78], Nixon et al. [208, 239], and others [131, 243]. The computer vision community considers soft biometric features as *describable visual attributes* useful for the representation of an image. For human identification, this representation can describe gender [260], ethnicity [261], accessories [35], clothing style [269], and facial-shapes [260]. See also [213, 251, 312].

Anthropometric attributes have been used to measure the geometry, and shape of the face, body, and skeleton. These soft biometric traits have become important as middle-level features in some applications: from human identification to gender, ethnicity, and age estimation to emotion or expression recognition, and others. Adjeroh et al. [3], Cao [49], and Lucas [188] show that the correlation of the body and face measurements can be successfully used to predict some unknown body measurements, including weight, and to successfully discriminate duplicates in a dataset. These findings are quite important as they imply we can obtain good identification performance with only some anthropometric measurements. However, these results are based on handmade measures of the body, a highly constrained scenario, difficult to obtain in a surveillance setting. As noted in [77], in practical applications these methods have to account for several factors: correlation of the geometric measurements [3, 106], variations in sensor and calibration [72, 310], and fusion of the information from different sources (e.g., in multi-view systems).

5.1.1 Anthropometric Features From the Body

Natural Image Techniques

Among the many anthropometric measurements, the *body height* is the most prominent and easy to acquire. However, different challenges remain, including the *human pose*, which constitutes a primary nuisance. A simple solution adopted by BenAbdelkader and Davis [22] is

to take the average of different body measurements over different poses. They consider shoulder breadth in addition to height for improving multi-target tracking across multiple cameras. Other related works include Criminisi et al., taking advantage of the well-known work on single view metrology [72], [73], Nguyen et al. [207] using a cross ratio technique in parallel with the vanishing point method, and Madden and Piccardi [47] from surveillance video. Velardo et al. [290], inspired by [264] on height estimation, proposed a model-based approach to study the correlation between weight and other common anthropometric measurements. The analysis was based on manual measurements on a barely sufficient set of natural images, but using the anthropometric measurements from the NHANES [56] dataset for training.

3D Techniques in Anthropometric Soft Biometric

The natural setting for the acquisition of anthropometric measurements is the 3D space. Previous work, principally in 2D, have used simple geometric rules to extract the real measure from pixel distances. However, measuring curved surfaces with the respective projection can quickliy lead to erroneous measurements. Recently, leveraging the introduction of cheap 3D acquisition devices, such as Microsoft Kinect [71], has made it possible to acquire geometric information, with low hardware requirements. The key advantage of using Microsoft Kinect is the availability of the body tracker [266], which can detect the pose assumed by the body and retrieve the skeleton. Velardo et al. [291] extended the weight prediction scheme introduced in [290] by extracting anthropometric measurements automatically using the skeletal joints from the body parts tracker [266]. The method showed good results, but the small RGB-D dataset limits the evaluation to a restricted number of body shapes (15 subjects) assuming the same pose. Recently, Madadi et al. [189] presented a method to extract soft biometric features using depth sensors, and the body parts tracking algorithm [266]. This method assumes a multi-parts labeled training dataset, and that the subject is aligned to the best model in the dataset. These

constraints, although familiar to many 3D matching frameworks, make this approach quite limited, and not scalable to a high number of poses.

Body Shape Modeling

Related to 3D soft-biometrics, but not exploited in previous works, is the massive work in body modeling and character generation. Weiss et al. [295] showed that the Microsoft Kinect could effectively reconstruct the 3D representation of a body in a less constrained environment. However, the joint optimization involved in the registration and fitting of the 4 point clouds to the body model makes the system extremely slow (40 min for one subject). Recently, Bogo et al. [30] introduced a new method, and a new body model using only a frontal view of the body. However, this method requires previously selected skeleton joints on the image. For a deep understanding of the human body shape (and thus the anthropometric measurements) under pose transformation, we need to consider the body composition, and soft-tissue deformations under pose variation. Classical model-based human character modeling in computer graphics is based on the "layered character construction" framework [58]. A skeleton drives soft-tissue motions including kinematic deformations and dynamics (e.g., muscle bulging), with the fat/tissue layer represented by a low-resolution mass-spring model. Many methods have been developed for controlling the dynamic simulation of general rigged models [52] 2002, [51] 2007, using finite element methods (FEM). These physically-based models are often based on material properties of human soft tissue [195],[10],[161]. Recently, data driven approaches are becoming more popular [6],[7],[115],[127],[61]. Pons-Moll et al. [227] extended the SCAPE model [7] to deformations due to dynamic movement of the body using a high-resolution 4D capture system. These models can be particularly useful, not just for the body representation, but also for the synthesis of new bodies. Impressive was the work by Sutton et al. [266] on real-time tracking of body parts. They trained a complex random forest classifier using only synthetic data, gener-

ated by rendering synthetic characters. However, in [266], the number of different bodies was limited to less than 30, but the body poses were augmented by "moving" the character through MoCap data [67]. Recently, Piccirilli et al. [222] generated an entire population of synthetic individuals to predict the Body Surface Area (BSA). Remarkable is the possibility to generate characters with different body compositions, in particular, different lean/fat ratios. In this work, we extend the technique to multiple poses and anthropometric measurements.

5.1.2 Anthropometric Datasets

One difficult challenge in studying pose variations in human anthropometry is the lack of suitable datasets. Common datasets used in geometry processing lack anthropometric measurements, and are typically limited regarding the number of subjects, and thus cannot capture the large variations in a human population. On the other hand, datasets used for anthropometric measurements lack a rich set of 3D data for different poses. Prior related work [290],[291],[189] have conducted a small in-house acquisition using available cheap 3D cameras, and manual measurements for a few individuals. A commonly used anthropometric dataset is the CAESAR dataset [244]. This is however expensive, with a limited number, and diversity of subjects (2400 subjects). These are now becoming inadequate for population-based study, especially, when demographic stratification is considered. Perhaps more importantly, the dataset did not consider the pose problem, and hence individuals were not measured under different poses. For a better understanding of human body shape and measurements under pose transformations, we need to have the anthropometric measurements under various poses. We are not aware of any human anthropometric dataset with this key information.

5.1.3 Main Contributions

We introduce four major contributions in the area of anthropometric soft biometrics. We describe these briefly below. **Analysis of pose variation in human anthropometry**. We present a detailed statistical analysis establishing the difficult challenge of using anthropometric measurements as soft-biometric traits under pose variations. Although the problem is intuitive, to our knowledge, this is the first detailed statistical analysis of the impact of the pose on human anthropometry in the biometrics literature.

Spectral geometry approach to soft biometrics. One crucial requirement for a modern surveillance system is the pose invariance. Pose variation can be formalized as a Euclidean transformation applied to the skeleton joints. However, due to the soft-tissue composition of the human body, the body shape gets deformed following nonlinear laws. Such non-linearity depends on various factors, such as age, gender, and body composition [74]. We present an SG approach to the traditional biometric tasks of identification, verification, and retrieval of body shape under pose variation. We use known local spectral descriptors capable of representing deformations of the body due to individual body morphology, but still able to be invariant to pose transformations. At the same time, we report the results obtained using the anthropometric measurements as geometric features, and comparing the two methods. To our knowledge, the present work constitutes the first attempt at using spectral geometry techniques for soft biometric description.

Semantic prediction via spectral geometry. Soft biometric, and anthropometric measures constitute semantically informative features since they carry important information about the body that can be described by humans [238](e.g., height: 5'2": short, waist circ. 45": chubby, etc.). Unfortunately, spectral descriptors are machine oriented and less human interpretable. To recover the semantic information we propose a predictor able to regress common geometric attributes from the spectral representation. We show prediction results of common global

measure, e.g., waist-to-height ratio, under pose deformations, a hard task for many automatic systems, as well as handmade measurement systems.

Virtual Pose Dataset (VPD). Given the lack of datasets for studying the impact of pose variations in human anthropometry, in this work, we introduce a new synthetic dataset composed of 3D body shapes of different individuals in various poses, along with their anthropometric measurements. We call this the Virtual Pose Dataset (VPD). Example poses from this dataset are shown in Fig. 5.1. This dataset constitutes the ideal setting to study body geometries along pose variations in a more controlled environment.

5.2 Variability of Anthropometric Measurements under Pose Transformations

Given the human body composition (bone, cartilage, and soft-tissue), and the numerous poses it can assume, there is a significant variation in the human body appearance, as well in the geometry, and thus in the anthropometric measurements. This behavior is known but has never been studied in an anthropometric soft-biometric setting. Traditional approaches often constrain the measurements to a few well known (canonical) poses and consider only measurement errors on the same pose. These are now becoming inadequate, especially for unconstrained environments, or uncooperative/deceptive individuals. We conduct a statistical analysis of the traditional anthropometric measurements to understand the variability of these measurements under pose transformation. We use three techniques: repeated measures, post-hoc analysis, and mixed effect analysis.

Repeated measures

Given a population W of N subjects $W = \{S_{1,\ldots,N}\}$ in a default pose T: $W_T = \{S_{jp} \quad j = 1,\ldots,N \quad p = T\}$, each anthropometric measurement g_i has a mean value, and variance, over the entire population:

$$\bar{g}_i(W_T) = \frac{1}{N}\sum_{j=1}^{N} g_i(S_{jT}) \tag{5.1}$$

$$\sigma^2(g_i(W_T)) = \frac{1}{N}\sum_{j=1}^{N}(g_i(S_{jT}) - \bar{g}_i(W_T))^2 \tag{5.2}$$

If each subject in the population W_T can assume P poses $S_{jp} \quad p = 1,\ldots,P$, we consider the mean value, and variance of the anthropometric measurement g_i for the individual S_j over the entire set of poses:

$$\bar{g}_i(S_j) = \frac{1}{P}\sum_{p=1}^{P} g_i(S_{jp}) \tag{5.3}$$

$$\sigma^2(g_i(S_j)) = \frac{1}{P}\sum_{p=1}^{P}(g_i(S_{jp}) - \bar{g}_i(S_{jp}))^2 \tag{5.4}$$

In this new framework, the variability introduced by the pose transformation will additionally degrade the subject discriminability from the anthropometric measurements. Traditional soft-biometric systems operate on subjects with the same pose, thus on a subset of the original set ($W_{p=T} \subset W$), ignoring the pose. We consider a repeated-measures design, where each participant provides the anthropometric measurements at multiple poses. In this scenario, the assumption on the model errors is different for variances present between subjects. In fact, the population of subjects at pose T is not independent of the population in other poses, since the population is composed of the same subjects, which vary under different poses. In literature, this situation has been called with different names: block design, multilevel modeling, and repeated-measure design [102]. We partition the variability attributable to the differences between groups

of poses (SS_{BG} SS : sum of squares) and variability within groups (SS_{WG}) exactly as we do in a between-subjects (independent) ANOVA. However, with a repeated measures ANOVA, as we are using the same subjects in each group (dependent condition), we can remove the variability due to the individual differences between subjects, referred to as (SS_M), from the within-groups variability (SS_{WG}). We treat each subject as a block (of poses), and we can calculate this variability as we do with any between-subjects factor. Now that we have removed the between-subjects variability, our new error (SS_R) only reflects individual variability to each condition.

$$SS_{Total} = SS_{BG} + SS_{WG} \tag{5.5}$$

$$SS_{WG} = SS_M + SS_R \tag{5.6}$$

Given the subjects in pose *P* W_P , and the anthropometric measurements $g_{i=1,\dots,l}(W_P)$. Let $\bar{g}_i(W_P)$ denote the mean over the *P*-th group of measurement i. Our goal is to test

$$H_0 : \bar{g}_i(W_1) = \bar{g}_i(W_2) = \cdots = \bar{g}_i(W_P) \tag{5.7}$$

hypothesis that the means for the measurement i of the P dependent groups of poses are equal. To verify we compute the test statistic *F*, rejecting the hypothesis if $F \geq f$, where *f* is the $1-\alpha$ quantile of an F-distribution. For more details about the method, we refer to Wilcox [297].

Post Hoc analysis

The above method is designed to verify the null hypothesis, but it does not specify how the groups differ and how much they differ. A common procedure is to do pairwise comparisons between the groups, (e.g., $H_1 : \bar{g}_i(S_{j1}) = \bar{g}_i(S_{j2}), \bar{g}_i(S_{j1}) = \bar{g}_i(S_{j3}), \dots, \bar{g}_i(S_{jP-1}) = \bar{g}_i(S_{jP})$.

However this procedure is complicated by the fact that the individual tests are not all independent. We develop our Post Hoc analysis using multiple comparison among dependent groups using Rom's method [246] for controlling the family-wise error (FWE). The sequentially rejective method from Rom [246] computes significance levels for each of the $C = (P^2 - P)/2$ tests, and rejecting all the test with ordering label less than the critical value. The algorithm will run until all the C hypotheses have been tested.

Mixed effects analysis

The above methods test the null hypothesis (H_0), and compare the different groups (H_1) respectively. However, we have not shown any result proving that the pose information can affect the anthropometric measurements in a regression framework. To show that, we use a multilevel model [268] approach, designed as a simple regression that allows for the errors to be dependent on each other (as the pose condition is repeated within each participant). This method is composed of a linear mixed-effects model.

$$y \sim x + \epsilon \tag{5.8}$$

A mixed-effects model consists of two parts, fixed effects (x) and random effects (ϵ). Fixed-effects terms are usually the conventional linear regression part, and the random effects are associated with individual experimental units drawn at random from a population. The random effects have prior distributions modeled as **random intercepts**, whereas fixed effects do not. Mixed-effects models can represent the covariance structure related to the grouping of data by associating the common random effects to observations that have the same level of a grouping variable [15].

Similar to any approach to model testing, we want to see if our predictive model, augmented

(with the pose) is better than a simple one parameter mean model. Thus, we specify a baseline model in which the anthropometric measurement g_i, is predicted by its overall mean $g_i \sim \bar{g}_i$. Second, we specify our model of interest, in which the anthropometric measurement g_i is predicted by the pose other than the mean $g_i \sim \bar{g}_i + pose$, which was repeated within subjects. Multiple responses (measurements) from the same subject (at different poses) cannot be regarded as independent of each other. Every person has a different body shape, and this is going to be an idiosyncratic factor that affects all responses from the same subject, thus making these different responses inter-dependent rather than independent. Adding a random effect for each subject allows us to resolve this non-independence by assuming a different "baseline" measurement value for each subject. For instance, within the male and the female groups, you see lots of individual variation, with some people having relatively higher values for their sex and others having relatively lower values. We can model these individual differences by assuming different **random intercepts** for each subject. In the mixed model, we add one or more random effects to our fixed effects. These random effects essentially give structure to the error term ϵ. In the case of our model here, we add a random effect for subject/pose, and this characterizes the idiosyncratic variation that is due to the grouping of the subjects by pose. In R style notation:

$$g_i = \bar{g}_i + pose + (1|subject/pose) + \epsilon \tag{5.9}$$

The general error term ϵ is necessary because even if we accounted for individual variation, there is still going to be "random" differences between the measurements of individual subjects. The results of this analysis on the VPD data set are presented in Section 5.4.

5.3 Spectral Geometry Approach to Soft Biometrics

In this section, we present our approach to addressing the problem of pose in whole-body soft biometrics. We start with the spectral features, the cornerstone of our approach.

Spectral features

For a given subject, we model the shape as a two-dimensional Riemannian manifold S sampled at n points $s_1, \ldots, s_n$ and represented as a triangular mesh. Let f be a real-valued function, with $f \in \mathcal{R}^2$, defined on S. A generalization of the Laplacian operator Δ for Riemannian manifold surfaces is:

$$\Delta_S f = div(\nabla(f)) = \frac{1}{\sqrt{det(g_{ij})}} \sum_{i.j} \partial_i (g^{ij} \sqrt{det(g_{ij})} \partial_j f) \tag{5.10}$$

with

$$g_{ij} = \begin{pmatrix} \langle \partial_i f, \partial_i f \rangle & \langle \partial_i f, \partial_j f \rangle \\ \langle \partial_j f, \partial_i f \rangle & \langle \partial_j f, \partial_j f \rangle \end{pmatrix} \tag{5.11}$$

the first fundamental form, $g^{ij} = (g_{ij})^{-1}$, and ∇:gradient.

Let $\mathcal{V}$ and $\mathcal{W}$ be vector spaces, and let $T : \mathcal{V} \to \mathcal{W}$ be an injective linear transformation. T is said to be isometric if for all $v, v' \in \mathcal{V}$

$$\langle T(v), T(v') \rangle_{\mathcal{W}} = \langle v, v' \rangle_{\mathcal{V}}, \tag{5.12}$$

and by definition keep the same surface area. As the Laplace-Beltrami operator (LBO) Δ_S depends only on g_{ij}, for each $u, v \in S$

$$\langle T(g_{ij}(u)), T(g_{ij}(v)) \rangle_{S'} = \langle g_{ij}(u), g_{ij}(v) \rangle_S. \tag{5.13}$$

Then $\Delta_S = \Delta_{S'}$ for any isometric function $TS \rightarrow S'$. Thus for body transformations that are isometries, the LBO can completely characterize the shape without loss of information. However the LBO depends to the surface area considered by the Laplacian operation ($det(g_{ij})$). Thus, in presence of more general transformations the LBO can become not invariant (e.g., scaling: $T : \alpha S \rightarrow S'$, $\alpha\Delta_S = \Delta_{S'}$). The LBO spectra is obtained as the solutions of the Helmotz equation $\Delta_s\phi_l = \lambda_l\phi_l$, denoted as $\{\phi_l, \lambda_l\}_{l\geq 1}$. In particular, the eigenvalues λ_l, assume the global descriptors of the shape. Interestingly, the eigenvalues λ_l are covariant with $det(g_{ij})$, thus with the surface area. This means that the body surface area is part of the LBO spectra, and theoretically can be used as an anthropometric feature. Local descriptors X, instead, try to represent the geometric structure within a small neighborhood of a point. For each sample location $s_i \in S$ we can compute a *q*-dimensional local descriptor $\mathbf{X}(s_i) = (X_1(s_i), \ldots, X_q(s_i))^T$. The use of local descriptors in this work is motivated by the local deformation of the shape with the body composition (e.g., lean/fat ratio), but invariant to global transformation.

HKS The Heat Kernel Signature (HKS) [276] is based on the heat diffusion process over the shape, governed by:

$$\frac{\partial S}{\partial t}(x,t) = \Delta S(x,t) \tag{5.14}$$

The heat kernel associated with ΔS is given by:

$$h_t(s_i, s_j) = \sum_{l\geq 1} e^{\lambda_l t}\phi_l(s_i)\phi_l(s_j) \tag{5.15}$$

where ϕ_l represents the *l*-th eigenfunction, λ_l the *l*-th eigenvalue of the LBO, and t the time intervals considered for the diffusion process. Sun et al. [276] proposed to use the diagonal of the heat kernel taken at *q* log-sampled time values ($t = \alpha^\tau, \tau = 1, \ldots, q$) as a local intrinsic

feature descriptor, called the heat kernel signature (HKS)

$$\mathbf{HKS}(s_i) = (h_{\alpha^{\tau_1}}(s_i, s_j), \ldots, h_{\alpha^{\tau_q}}(s_i, s_j))^T. \tag{5.16}$$

This formulation, however, is only invariant to isometric deformation, but not to scale transformations.

siHKS To make the HKS scale invariant, Kokkinos and Bronstein [40] developed a scale-covariant heat kernel

$$\hat{h}_\tau(s_i, s_i) = \frac{-\sum\limits_{l\geq 1} \lambda_l \alpha^\tau \log \alpha e^{-\lambda_l \alpha^\tau} \phi_l^2(s_i)}{\sum\limits_{l\geq 1} e^{-\lambda_l \alpha^\tau} \phi_l^2(s_i)} \tag{5.17}$$

that undergoes shift in τ by $2\log_\alpha c$ as a result of shape scaling by a factor of c. In the Fourier domain, this shift results in a complex phase $\hat{H}(\omega)e^{-i\omega 2\log_\alpha c}$, where $\hat{H}(\omega)$ denotes the Fourier transform of $\hat{h}_\tau$ w.r.t τ. The scale-invariant HKS (siHKS) is constructed by taking the absolute value of $H(\omega)$ and sampling $|H(\omega)|$ at *q* frequencies:

$$\mathbf{siHKS}(s_i) = (|H(\omega_1), \ldots, H(\omega_q)|)^T \tag{5.18}$$

WKS The wave kernel signature (WKS) [11] follows a similar idea to the HKS, replacing the heat equation with the Schrödinger wave equation. Assuming that the Laplace spectrum of the shape has no repeated eigenvalues, the wave function of the particle is given by:

$$\Psi_E(x, t) = \sum_{k=0}^{\infty} e^{iE_k t} \phi_k(x) f_E(E_k) \tag{5.19}$$

where $f_E^2(E_k)$ is the energy probability distribution with expectation value E. The probability of measuring the particle at a point $x \in S$ is then $|\Psi_E(x, t)|^2$. Aubry et al. [11] define the WKS

as the average probability, over time, to measure a particle in x:

$$\mathbf{WKS}(E, x) = \sum_{k=0}^{\infty} \phi_k(x)^2 f_E(E_k)^2 \tag{5.20}$$

Thus the energy is directly related to the LBO eigenvalues, and therefore to an intrinsic notion of scale in the shape.

5.3.1 Spectral and Anthropometric Matching

To evaluate the performance of the defined local spectral descriptors against the traditional anthropometric measurements, we devise a biometric scenario, where the descriptors with a higher degree of pose invariance will have a high similarity value for the same subject under different poses and lower values with all other subjects in whatever pose. Given a set of local *q*-dimensional spectral descriptors (e.g.,HKS, siHKS, WKS) computed without loss of generality at *n* sampled points of the shape (mesh vertices), we represent them as a $q \times n$ matrix $\mathbf{X}$

$$\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n) = (\mathbf{x}(s_1), \ldots, \mathbf{x}(s_n)). \quad \mathbf{x} = \{x_1, \ldots, x_q\}^T \tag{5.21}$$

The matrix $\mathbf{X}$ represents a dense description of the subject's shape. A simple distance metric can be used by computing the Frobenius norm of the respective matrices $||\mathbf{X}_i - \mathbf{X}_j||_F$. Similar to the spectral matrix, we construct an anthropometric descriptor for the shape S as the set of *a* anthropometric measurements $(g_i(S), i = 1, \ldots, a)$. Each measurement is given by the distances of the mesh vertices that constitute the minimum path on the body surface $g_i = \sum_l ||s_l - s_{l+1}||_2, l = \{s_i \in P(g_i)\}$ (see Fig. 5.2). Each path P is constituted by a set of different number of points, thus the matrix $\mathbf{M}$ of sampled anthropometric measurements is not square. For simplicity we just consider the vector $\mathbf{m}_i$ of anthropometric measurements as the

Figure 5.2: Some anthropometric measurements using the MakeHuman (MH) mesh model.

anthropometric features:

$$\mathbf{m} = (g_1(S), \ldots, g_a(S)). \tag{5.22}$$

With this formulation, the similarity metrics for two shapes S_i and S_j is the L_2 norm $||\mathbf{m}_i - \mathbf{m}_j||_2$ of the anthropometric vectors. We devise the follow experiment: for each subject' shape S_j, we consider a family $F(j) = \{S_{ji}, \quad i = 1, \ldots, k\}$ of shapes composed of the original and k shapes obtained with a pose transformation. For each $S_{ji} \in F(j), j = 1, \ldots, n$ in the population W, we compute the anthropometric measurements, together with the spectral descriptors. A simple classifier, using the above metrics, will try to classify the subjects as a genuine (same subject, but different poses), or as an impostor member (different subjects). From a biometric standpoint, the problem is highly challenging because the body soft-tissue is subject to nonlinear deformations, affecting the shape and thus the respective anthropometric measurements.

5.3.2 Soft Biometrics from Spectral Features

The spectral representation $\mathbf{X}(S)$ of the shape S has been previously used in 3D shape retrieval [215],[241],[171],[165],[223]. However, these methods "retrieve" the semantic informa-

tion of the shape (lean, fat, similar, cat, dog, etc.) by comparing the given shape with other well known (canonical) shapes. The major reason is due to the lack of semantic information as meta-data in the dataset. Using the proposed framework, where are available anthropometric measurements as well as spectral descriptors, we can learn semantically meaningful representations. In the present framework, we identify the semantic information as anthropometric soft-biometrics, not necessarily in the form of anthropometric measurements. In fact, more meaningful indicators, such as those used in medicine, corresponding to a combination of two or more anthropometric measurements can be used as soft-biometrics. A well-known indicator is the body mass index (BMI) [29]. Although computed from stature and weight, we do not consider here since it is not a geometric information. Two examples of these "combination indicators" are the recently introduced ABSI (A Body Shape Index [151]), and SBSI (Surface-based Body Shape Index [234]), both of which were shown to outperform the BMI in mortality prediction. Another index is the whole body surface area (WBSA) [83]. This quantity, other than providing useful medical information, it assumes a principal role in spectral geometry as seen before. Interestingly, the WBSA can be predicted with computer vision techniques, even when the body is partially visible [222]. In this work, we focus on the waist-to-height ratio (WHR or WtHr) as a global semantic attribute for two major reasons. The WHR is known to be a more valid health indicator than BMI [160],[41],[9]. The WHR, as a shape index, easily captures the body shape appearance (slender athletic vs. obese) using just anthropometric measurements.

We devise a supervised framework able to predict a semantic characterization of the body shape from its spectral description. Given the pairs $\{\mathbf{X}(S_i), y_i\}_{i=1}^{L}$ composed of the subject's spectral shape representation $\mathbf{X}(S)$, and the corresponding shape semantic value y, the system will be able to retrieve the soft biometric information from the shape representation $\mathbf{X}(S)$ under pose variation.

Encoding

Given the high number of samples from the surface (thousands), with low discriminative power, an essential step is to encode the spectral features in a compact and richer representation. This problem has driven different works from machine learning (feature encoding [38], dictionary learning [177]) and information theory [176] to non-rigid shape representation. A common computer vision solution is the use of encoding techniques such as the Fisher vector (**FV**) [217],[218]. The **FV** is an encoding of the features obtained by pooling local descriptors. It is frequently used as a global image descriptor in visual classification. Let $X = \{x_t, t = 1, \ldots, T\}$ be a set of d-dimensional local descriptors (e.g., SIFT, HKS, WKS). Assuming independent samples (assumption relaxed in the normalization step [218]), the Fisher vector $\mathscr{G}$ of shape S is given by:

$$\mathscr{G}_\lambda(X) = \frac{1}{T} \sum_{t=1}^{T} L_\lambda \nabla_\lambda \log \mu_\lambda(x_t) \tag{5.23}$$

as the sum of normalized gradient statistics $L_\lambda \nabla_\lambda \log \mu_\lambda(x_t)$. Here, u_λ is a Gaussian mixture model (GMM) with diagonal covariance of parameter λ, which models the generation process of local descriptors, traditionally called universal (probabilistic) visual vocabulary [217]. L_λ is the Cholesky decomposition of the Fisher information matrix of u_λ. The **FV** encoding creates an embedding of the local descriptors x_t in higher-dimensional space, which is more amenable to linear classification. Moreover, each subject's spectral representation $\mathbf{X}(S) \in \mathcal{R}^{q \times n}$ is encoded in a single vector, representing the deviation from a "universal" generative model, learned offline from a large set of samples. This characterization is given as a gradient vector w.r.t. the parameters of the model (λ). The **FV** $\in \mathcal{R}^{2 \times k \times d}$, where *k* is the number of atoms in the visual vocabulary, and *d* the dimensionality of the descriptors.

Semantic Feature Prediction

Given the pairs $\{\boldsymbol{\varphi}_i, y_i\}_{i=1}^{L}$, where $\boldsymbol{\varphi} = \mathscr{G}_\lambda(X(s_k))$ is the Fisher vector of $X(S)$, and $y \in \mathcal{R}$ a numerical value representing the semantic feature to predict (e.g. Waist-to-Height-Ratio **WHR**), our goal is to learn a mapping: $\boldsymbol{\varphi} \rightarrow y$ able to predict the subject's semantic information. We parametrize the transformation as a regression function $T(\boldsymbol{\varphi}_i, \boldsymbol{\Theta})$ such that $y_i \approx T(\boldsymbol{\varphi}_i, \boldsymbol{\Theta})$, with $\boldsymbol{\Theta}$ the regressor parameters. We minimize the mean square error as in a typical regression task:

$$L(\Theta) = \sum_{j=1}^{J} \frac{1}{2} ||y_i - T(\boldsymbol{\varphi}_i, \boldsymbol{\Theta})||_2^2 \tag{5.24}$$

We use a multi-layer feed-forward neural network (see Fig. 5.3) as high capacity regressor $T(\boldsymbol{\varphi}_i, \boldsymbol{\Theta})$. The network will learn a relation between the Fisher encoding of the spectral features, and the semantic feature (WHR). With this framework, it is theoretically possible to regress almost all geometric quantities (including anthropometric measurements) of the body from the spectral description. However, there are some limitations in the representation power of the spectral descriptors. Reuter proved that it is possible to create continuous families of manifolds with the same spectrum (**isospectral**), which does not entirely determine the object geometry [240]. Although, the compactness theorem [26] shows that the spectrum does place some strong constraints on the geometries allowed by a given spectrum.

5.4 Results

In this section, we present the datasets and the results of the proposed methods. For the spectral descriptors, we used the methods and codes from [11, 38] in Matlab. The statistical analysis of the anthropometric features was done in R [232]. The regression framework was implemented with Keras [64], and Tensorflow [1].

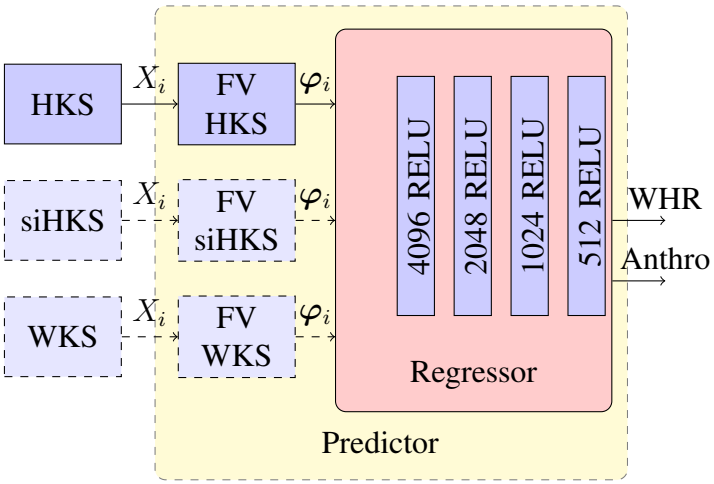


Figure 5.3: Anthropometric soft-biometrics predictor.

5.4.1 Datasets

In this section we present a new data framework.

Virtual Pose Dataset (VPD)

We designed a new synthetic dataset composed of 3D body shapes from different individuals assuming various poses. We call this the *Virtual Pose Dataset (VPD)*. Building on Piccirilli et al. [222] work, we used the MakeHuman (MH) tool [16] to create different human characters with their 3D mesh and relative anthropometric measurements. MH uses a layered character construction framework [58], where the subject body dimensions can be decided, and measured via manuals controls. The presented framework is general, and can be easily upgraded with more elaborate body models, for instance [227],[30],[31]. The huge advantage in using [222] is the automatic generation of thousands of subjects, without manual intervention. Using the well known NHANES [56] dataset, we can automatically replicate subjects with given anthropometric dimensions in the 3D mesh. The novelty is the extension of [222] to multiple poses (Fig. 5.1), with the possibility to save the anthropometric measurements automatically for each

pose (see sample measurements in Fig. 5.2). In the present work, we do not use clothes on the synthetic characters, although present in MH, since the focus is on the soft-tissue deformation under pose variations. Moreover, since the dataset is used to benchmark the anthropometric measurements under pose variations, the clothing component just an additional layer on the 3D layered skinned model, thus adding a "noisy" component to the measures. Although there have been different attempts to body analysis under clothing [46],[116],[65], we refer to them for more information. MH body model is a triangular mesh composed of 14444 vertices (samples), and 28796 faces. Its deformation engine permits to obtain bodies with different measures, and poses, with limited mesh artifacts.

The VPD contains 132 subjects (66 males, 66 females) with anthropometric measurements. For each subject, we selected 18 different poses (Fig. 5.1), for a total of 2376 total meshes. We decided to select poses from different groups: standing, gym, sit, and some more unlikely: fly, and fight, to cover a broad range of variations for a more thorough analysis.

For every subject, MH computes 19 measurements, namely, WBSA [83], height, hips circumference, waist circumference, bust circumference, under-bust circumference, neck circumference, front chest, upper arm length, upper arm circumference, lower arm length, wrist circumference, shoulder distance, upper leg height, thigh circumference, calf circumference, lower leg circumference, ankle circumference, and knee circumference. The measures are based on the geodesic distance on the body surface, similar to a measuring tape. Bulging, and swelling of soft-tissues with the pose will inevitably affect these measures. For the stature, *being a measure independent of the body composition*, we could just report the ground truth at the default pose T. However, measuring the stature of different poses can be a daunting task. Using geodesic methods lead to the same soft-tissue deformation problem. An attractive solution is to use the lengths of the skeleton bones. Although Kinect body tracker [71] offers good results, the accuracy of the measurement is heavily dependent on the tracker performance. We decided to limit

our analysis to two simple methods: surrounding bounding box of the body (Height BB), and toe to head Euclidean distance (Height HT).

FAUST

To evaluate the performance of the developed method on acquired real data, we used the FAUST dataset [31]. FAUST contains data on ten subjects. The subject meshes are obtained from high-speed acquisitions of moving subjects, containing more natural deformations. The mesh models have missing parts caused by occlusion, and topological noise where touching body parts are fused together, or just hidden (e.g., under feet, armpit). The dataset also contains some non-manifold vertices and edges, which some retrieval methods cannot handle. We, therefore, used a version of the data from which these non-manifold components were pre-processed, creating a watertight manifold for each model, as specified in the FAUST challenge [31]. However, these reconstructed areas still affect the total surface area, thus the descriptors. Unfortunately, the FAUST dataset does not provide anthropometric measurements. We report the results only for spectral feature evaluation.

5.4.2 Anthropometric Measurements – Impact of Pose

In Fig. 5.4 we report some statistics of the anthropometric measurements for the subjects in standard T-pose for the VPD dataset. More extensive information is available in the supplementary material.

Repeated Measures ANOVA

We execute one way repeated measure analysis using the Wilcox robust estimation and testing package [190]. In this setting, each anthropometric measurement is analyzed in comparison with multiple dependent trimmed groups (10% of trimming), see Table 5.1. The test statistic

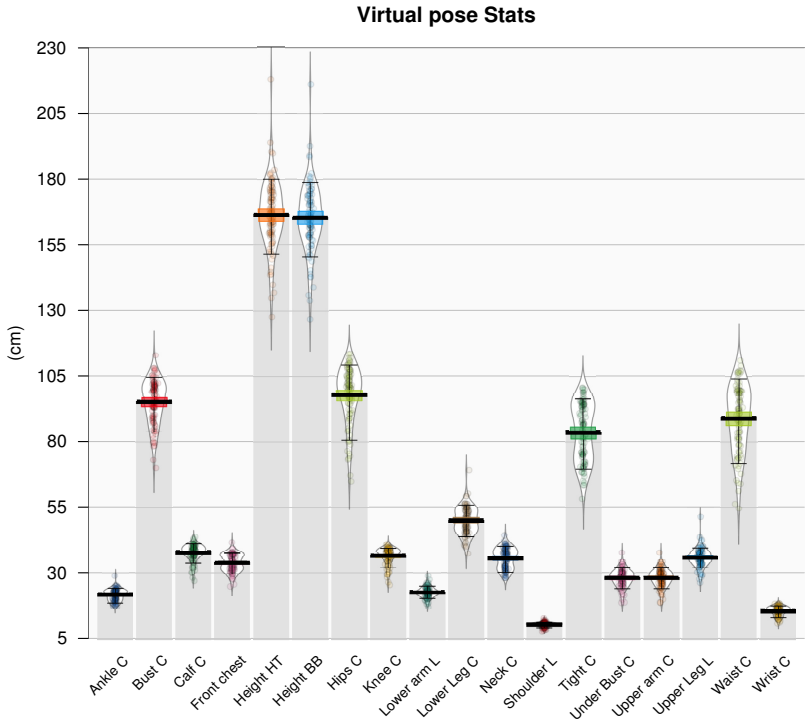


Figure 5.4: Statistics of anthropometric measurements for the T pose over subjects in the VPD. Points represent raw data, vertical bar indicates central tendencies, bean represents a smoothed density, colored rectangle denotes highest density interval quantities.

presents high values for the majority of the body measurements (test $>> 1$). However, as noted in [297], this does not mean that the null hypothesis is invalid for all the measurements. In fact, the explanatory measure of effect size $\hat{\xi}$ ranges from small to medium effect size. For an improved understanding of the behavior of the different anthropometric measurements, we perform post-hoc tests on different groups of poses.

Comparing Dependent Groups

Post-hoc tests consist of pairwise comparisons of all of the different combinations of the group means. We take every pair of groups and perform a t-test on each pair. The price paid for doing lots of tests is that each test is corrected to make it stricter so that across all tests the error rate is

Table 5.1: Repeated Measurement Anova results for some measurements.

Measure	$test$[1]	$\hat{\xi}$[2]	$df1$[3]	$df2$[3]
Waist	762.31	0.08	1.43	149.71
Calf	20372.52	0.06	1.00	105.47
Ankle	451.89	0.06	1.02	107.53
Bust	1131.03	0.07	1.20	125.87
Ubust	603.72	0.07	1.17	123.28
Neck	5448.16	0.13	2.20	230.73
Hips	896.90	0.10	1.66	174.46
Knee	22322.01	0.06	1.06	111.66
Thigh	3086.26	0.10	1.69	177.42
Wrist	7311.34	0.06	1.06	110.98
Uarmcirc	5511.86	0.12	2.07	217.34
Uarmlen	2464.09	0.13	2.28	239.13
lowleg	2310.32	0.06	1.05	110.76
Fchest	6743.97	0.11	1.83	192.23
WHR	6574.10	0.07	1.26	132.04
Hbb	18113.94	0.07	1.26	132.31
Htoe	13572.36	0.09	1.54	161.84
BSA	1135.97	0.07	1.22	128.31

[1] F-test statistic ; [2] Measure of effect size;
[3] degree of freedom;

controlled. We use the function `rmmcp` from the package [190] on trimmed means. Table 5.2 shows some of the key results (more in Supplementary Material). For instance, the waist circumference for poses 2 and 6: default vs. fight4 (see Fig. 5.1), presents a high value for the test statistic. Thus the waist circumference is affected by the pose transformation. However, for the same groups, other measurements are not affected, e.g., bust and calf circumference. In general, different measurements are affected differently by pose transformations (see Table 5.2, and Supplementary Material). Limb measurements covariant with the fat percentage are more prone to variations. For some groups of poses (e.g., standing), the pose transformation is limited to some body parts. Then only some anthropometric measurements may be affected. However, it is hard

to quantify the effect of the transformation, since the dynamic motion has a nonlinear effect on the soft tissues. In conclusion, the null hypothesis of the anthropometric measurements under pose transformation is often violated. Moreover, comparing different groups of poses shows an unpredictable behavior of the anthropometric measurements under pose transformation. This constitutes the main problem when attempting to use traditional regression methods since sparse outliers can drastically impact the overall performance.

Table 5.2: Some results of the Post-hoc analysis for comparing dependent groups on 10% trimmed means.

Measure	Gr1 vs Gr2	test[1]	p.value	p.crit	$\hat{\psi}^2$	ci lower[3]	ci upper[3]
Waist	2 vs 6	154	0.00	0.0009	0.9498	0.9267	0.9730
Calf	2 vs 6	-2.65	0.01	0.00	0.00	0.00	0.00
Bust	2 vs 6	-0.52	0.60	0.02	-0.01	-0.06	0.05
Neck	2 vs 6	93.66	0.00	0.00	1.72	1.65	1.79
Hips	2 vs 6	-42.22	0.00	0.00	-5.25	-5.72	-4.79
Tight	2 vs 6	121.74	0.00	0.00	0.24	0.24	0.25
Ankle	11 vs 13	-8.07	0.00	0.00	-0.04	-0.18	-0.07
Ankle	11 vs 14	-3.14	0.00	0.00	-0.04	-0.11	0.01
Ankle	11 vs 15	-4.48	0.00	0.00	-0.04	-0.13	-0.01
Ankle	11 vs 16	-1.67	0.10	0.00	-0.04	-0.09	0.03
Ankle	11 vs 17	4.22	0.00	0.00	0.15	0.01	0.13
Ankle	11 vs 18	-6.72	0.00	0.00	-0.04	-0.16	-0.05
Waist	10 vs 18	-38	0.00	0.0003	-1.7725	-1.9474	-1.5976
Waist	4 vs 7	-0.45	0.65	0.0169	-0.0007	-0.0068	0.0053

[1] T-test; [2] value of the test statistics; [3] confidence intervals;

Multilevel Analysis

We used R [232] and the `nlme` [225] package to perform a mixed effects analysis of the impact of the pose on the anthropometric measurements g_i. For the fixed effects term we consider the simple mean value (without interaction term), and the grouping effect of the pose as random effect term. In Table 5.3 we report the results of this analysis for some anthropometric measure-

Table 5.3: Linear mixed-effects model fit.

Measure	Model	AIC[1]	BIC[2]	logLik[3]	L.Ratio[4]
Waist	baseline	5157.30	5180.39	-2574.65	
	posemodel	7.31	128.55	17.34	5183.99
Hips	baseline	10286.05	10309.14	-5139.03	
	posemodel	4795.55	4916.79	-2376.78	5524.50
Ubust	baseline	5401.40	5424.49	-2696.70	
	posemodel	1020.70	1141.94	-489.35	4414.69
Bust	baseline	8486.49	8509.58	-4239.24	
	posemodel	2986.22	3107.45	-1472.11	5534.27
Thigh	baseline	590.68	613.78	-291.34	
	posemodel	-7158.37	-7037.14	3600.19	7783.06
Wrist	baseline	4511.65	4534.74	-2251.82	
	posemodel	-4959.02	-4837.78	2500.51	9504.67
Knee	baseline6	9943.26	9966.35	-4967.63	
	posemodel	-1370.83	-1249.59	706.42	11348.09
Neck	baseline	4850.44	4873.53	-2421.22	
	posemodel	-4130.85	-4009.61	2086.42	9015.29
UArm	baseline	3934.71	3957.81	-1963.36	
	posemodel	-4656.24	-4535.00	2349.12	8624.95
An-	baseline	1308.94	1332.03	-650.47	
kle	posemodel	-2784.76	-2663.53	1413.38	4127.70
Calf	baseline	-3775.55	-3752.45	1891.77	
	posemodel	-15401.41	-15280.17	7721.70	11659.86
Lleg	baseline	12854.44	12877.54	-6423.22	
	posemodel	6101.80	6223.03	-3029.90	6786.65
ULeg	baseline	9428.50	9451.59	-4710.25	
Ht	posemodel	2590.62	2711.86	-1274.31	6871.88

[1] Akaike Information Criterion; [2] Bayesian Information Criterion;
[3] Log-likelihood; [4] Log-likelihood Ratio;

ments with some important model fit indicators. Generally, with AIC (i.e., Akaike information criterion) and BIC (i.e., Bayesian information criterion), lower values indicate a better model, as it implies either a more parsimonious model, a better fit, or both. The log likelihood ratio assumes values from 4128 to 11659, verifying our assertion about the dependence of the anthropometric measurements on the pose. From these results, it is evident that a traditional regression framework [3, 290, 291] should include the pose information to avoid a performance

loss in the prediction. However, estimating the pose transformation is not easy, and adds more complexity to the system.

5.4.3 Spectral Features for Soft Biometrics

In this section, we compare the discriminative power of the spectral features against traditional anthropometric measurements. We use the spectral descriptors (HKS, siHKS, WKS), and the anthropometric vector on the VPD dataset, and the spectral descriptors on the FAUST dataset [31]. We pre-process the two datasets to extract the spectral features, while the anthropometric measurements were given by the VPD framework. We used the same pipeline as in [11],[38]: we compute the LBO eigenvalues from the body mesh with the cotangent formula [226], and subsequently the local spectral descriptors: HKS, siHKS, WKS. We consider 300 eigenvalues λ_l. The HKS has been computed for 23 time intervals $t_i = 2^\tau, \tau = \{5, 5.5, \ldots, 16\}$, the siHKS for 19 scaling factors $\omega = \{2, \ldots, 20\}$, and the WKS for 20 time intervals. We evaluate the performances on the task of verification, identification, and retrieval. We note that soft-biometrics are seldom used independently for these tasks, given their low performances. However, Lucas [188] showed that it is possible to come close to the identification rate of hard biometrics systems (fingerprints, iris), using body and face measurements. Lucas' work, although interesting, cannot be compared with the present framework since we do not consider facial measurements, and the method in [188] does not consider pose variations.

For the verification task, we compute the ROC curve in Fig. 5.5 using the round robin method [170]. We consider 18 samples for each subject in VPD dataset, for a total of 20,196 genuine, and 2,801,304 impostor scores. While ten samples for each of the ten subjects in the Faust dataset, for 4,500 genuine, and 495,000 impostor scores. In the VPD dataset, the HKS obtained the best performance with Area Under Curve (AUC)=0.99, then siHKS AUC=0.91 and the WKS with AUC=0.61. The anthropometric descriptor obtained AUC=0.66.

We evaluate the closed-set identification performance computing the cumulative match characteristic (Fig. 5.6). We used the same round-robin method to compute the identification rate. In the VPD dataset the HKS is still the best descriptor with an identification rate at rank-1 of 98.66 %, then siHKS with 12.34 %, and the WKS with 6.55 %. Using the Faust dataset we obtain different results, with WKS at 84.44 %, HKS 65.55 %, and siHKS 56.66 %. Although, the results on the Faust dataset are less reliable given the weakness of the CMC curve on small gallery size. The identification rate for the anthropometric vector is 5.16 %, the lowest of the tested descriptors.

We evaluate the performance in the retrieval task of recovering the subject independently by the pose using the precision-recall curve (Fig. 5.7). For this experiment, we consider relevant the subjects with the same identity. On the VPD dataset, the HKS produced the best performance, followed by the siHKS, and WKS, that still underperform the anthropometric vector. Interesting how, for low recall rate the anthropometric vector obtain 80 % precision. However, for increasing recall rate the precision rapidly decrease to under 10 %. Results for spectral features decrease less rapidly, performing better for high recall rates. For the FAUST dataset, the behavior is similar to previous observations, Table 5.4 reports some significant indicators to understand the performance.

The HKS largely outperform the traditional anthropometric features. However, we observe an unexpected behavior: siHKS and WKS underperform the HKS, also by a large margin, with WKS underperforming even the anthropometric vector. This behavior contradicts previous results in non-rigid shape retrieval, where newly developed descriptors (siHKS, WKS), allowing larger families of invariances, like scale, and topological transformations, reach the state of the art. In our setting, with a large population of bodies, the scale invariance property makes the spectral features of similar subjects but differing by a scale factor, identical. As a consequence, the scale invariance makes the surface area, fundamental information in the LBO spectrum, in-

sensible to different subjects. Unfortunately, while it is a suitable property for shape retrieval, where the goal is the retrieval of the correct shape (e.g., kids and adults are in the same category: human), it is unacceptable in anthropometric soft-biometric, where the principal task is the discrimination of body geometries. WKS, which allows a more substantial degree of invariance, and also attenuating lower frequencies of the spectrum, where global information are stored, perform poorly on the challenging VPD, but usually is state of the art in shape retrieval tasks. Although, the scenario changes in the presence of real data. We can see on the FAUST dataset, where there is more reconstruction noise, the HKS advantage, due to the limited surface area variation under isometric transformation, is drastically attenuated, limiting the descriptor performance. WKS, and siHKS being more robust, can outperform the HKS on more challenging data. This result constitutes a remarkable novelty, giving interesting information for future work.

Table 5.4: F-measure and D-prime.

Descr.	F-measure	D-prime
VPD HKS	-0.5933	0.8402
VPD siHKS	-0.6093	1.0979
VPD WKS	-0.2106	0.4686
VPD Anthro	-0.2346	0.4686
FAUST HKS	-1.3286	2.0219
FAUST siHKS	-07956	1.4643
FAUST WKS	-0.8117	1.6211

5.4.4 Predicting Semantic Features

The goal of the proposed prediction framework is to regress some geometric information of the body independently of the pose using the spectral features. For this experiment, we use the waist-to-height ratio (WHR) in the T-pose as a reference value. WHR is an important indica-

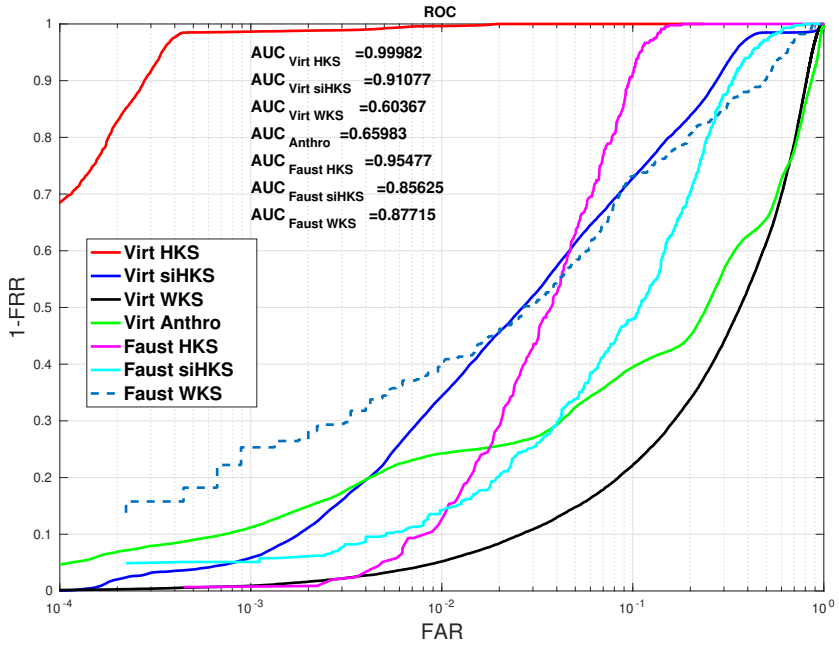


Figure 5.5: Receiver operating characteristics for the L_2 classifier based on spectral, and Anthropometric features for the Virtual and FAUST datasets.

tor since it can describe the body appearance at a distance (fat, slender, short, tall). For each mesh in the dataset S_i, we compute the Fisher vector $\boldsymbol{\varphi}_i$ for different dictionary dimensions (4,16,64). The feed-forward network $T(\boldsymbol{\varphi}_i, \boldsymbol{\Theta})$ consists of 4 fully-connected layers with rectified linear unit (ReLu) activation function [204] (see Fig. 5.3). The network hyperparameters $\boldsymbol{\Theta}$ are optimized using RMSprop [284], with MSE loss function, for 60 epochs. We evaluated the method over all the 2376 subjects using K-fold cross validation with K=10 (9/10 training, 1/10 test). Table 5.5 shows the K-fold cross-validation results for different dictionary sizes. Interestingly, the siHKS descriptor produced the best overall performance, followed closely by HKS, and WKS. High dictionary dimensions are not required for the spectral descriptors (see Table 5.5), due to the low descriptive power of the shape features. Typical WHR values are in the range of (0.3$\sim$0.7). Thus, the prediction error using siHKS accounts for 1/10th of the values, which is acceptable for both medical and soft-biometric applications.

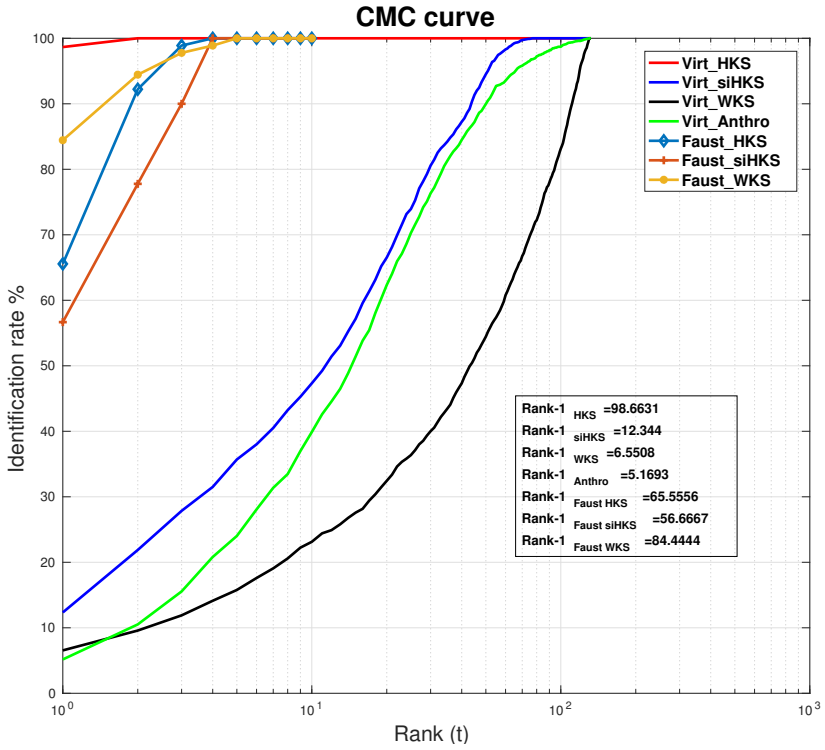


Figure 5.6: CMC for the L_2 classifier based on spectral, and Anthropometric features for the Virtual and FAUST datasets.

Table 5.5: WHR regression results. K=10 Fold cross validation.

Descr	Dict.[1]	MSE±std[2]	MAE±std[3]
HKS	64	0.00097±0.00076	0.0224±0.009
HKS	16	0.00092±0.00058	0.0158±0.002
HKS	4	0.00107±0.00037	0.0252±0.008
siHKS	64	0.00014±0.00014	0.0116±0.004
siHKS	16	0.00031±0.00023	0.0108±0.005
siHKS	4	0.00028±0.00024	0.0117±0.005
WKS	64	0.00215±0.00050	0.0349±0.005
WKS	16	0.00224±0.00047	0.0379±0.004
WKS	4	0.00273±0.00043	0.0414±0.003

[1] GMM dimension;
[2] Mean Squared Error with Standard deviation;
[3] Mean Absolute Error with Standard deviation;

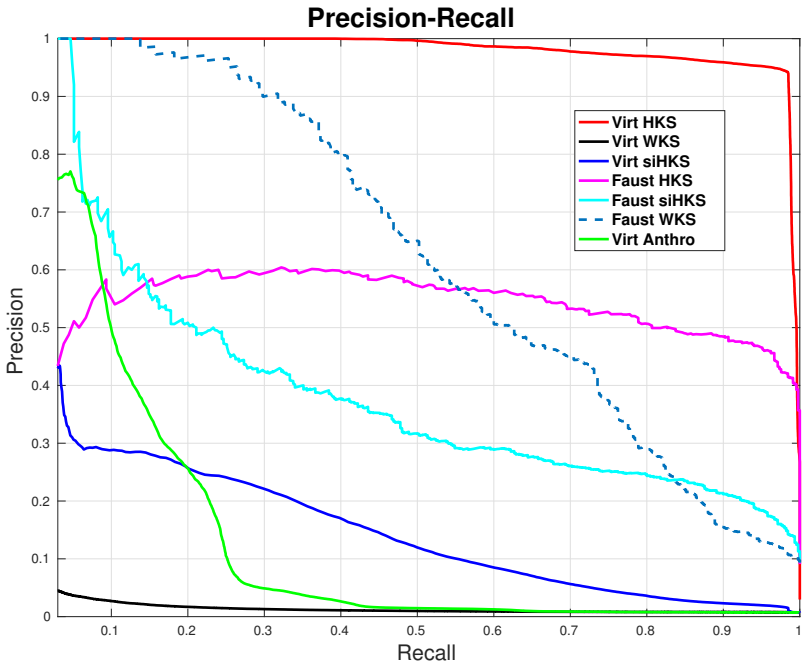


Figure 5.7: Precision-Recall for the L_2 classifier based on spectral, and Anthropometric features for the VPD and FAUST datasets.

5.5 Conclusion

Anthropometric soft-biometrics is an emerging field that is gaining more attention with the introduction of more powerful and efficient computer vision techniques. Leveraging these methods, we present an innovative framework to study the variability of human anthropometric measurements under pose transformations. Prior works on this topic have been heavily limited due to the lack of detailed data with pose information. We propose a virtual solution that circumvents the expensive burden of data acquisition. Using recent results in human body modeling, we can reproduce soft tissue deformations that profoundly affect the human body shape, and thus the anthropometric measurements. We present the Virtual Pose Dataset (VPD), a new dataset with 3D body models from multiple subjects under different poses. We show the inefficiency of traditional anthropometric measurements under pose deformation. To address the pose problem, we introduce a spectral geometry approach to anthropometric soft-biometrics,

defined as the geometric description of the human body. Our work is closely related to efforts in non-rigid shape retrieval. However, there are significant differences. Our notion of semantics is different from a similarity measure, connecting concepts from soft-biometric, medical indicators, and body modeling. We exploit these differences to propose a novel method for predicting body shape semantics based on the spectral geometric description of the human body. In doing so, we present an interesting application of the spectral description in learning useful medical, and soft-biometric quantities, namely, the combination body shape indices (e.g., WHR, WBSA) under pose variations, a task that has never been attempted in the literature. Experimental results on both our newly introduced virtual (VPD) and FAUST datasets with limited real data demonstrate the superiority of the spectral geometry approach to anthropometric soft biometrics. Future works will be focusing on testing more realistic scenarios, with only a portion of the body available as point clouds. This will be the most recurrent data in future surveillance systems, acquired from mobile or fixed lidar devices, RGB-D sensors, or simply as multi-view stereo reconstruction.

Chapter 6

Exploring the Human Body Manifold

In Chapter 4 (Figures 4.7-4.8) we introduced a graphical representation for a family of bodies. We also related this representation to known Probabilistic Graphical Models (PGM: CRF, MRF) [150]. The principal interest was the possibility to predict new bodies with semantic characteristics related to the family components. In this chapter, we generalize this idea to directed graphical models such as neural networks, and in particular, the new deep generative models [104],[103].

Let's suppose we have two non-similar subjects, given a similarity measure (e.g., waist circumference, stature, gender, WHR, etc.). We can ask a biometric system (e.g., the method proposed in section 4.5.3) to infer all the body variations occurring between these two subjects as new bodies. These new bodies are not part of the original dataset, however they need to be drawn from the same data-generating distribution. The system has to learn this distribution efficiently, as a low dimensional manifold embedded in a higher dimensional space, where the data reside. Discovering the structure of this space permits the analysis of body attributes, and the variations through different categories: male, female, adults, kids, etc., allowing the design of better machine learning systems for identification, and verification, as well as to regress

critical medical indices.

Assuming the space of bodies follow the manifold hypothesis [55],[263],[249]: natural data (images, videos, speech, etc.) are clustered in low dimensional compact manifolds with high prior, we can explore this domain by sampling, finding new bodies and evaluating those new results.

Given the problematic evaluation process of generated data: new bodies not present in training and test sets have no labels. We design a new regressor network, able to infer body characteristics, such as the WHR, from the unseen bodies. This method is unique, being able to annotate body proportions from generated data.

After an initial introduction to human body representation learning (Sections 6.1, 6.2), we describe recent deep generative models. Then we formalize the body generation problem (Section 6.3), as well as the evaluation method, and subsequently, we present the relative results. In this work, we use a rendered version of the VirtualBody dataset (Chapter 2). The new dataset creation is described in Appendix A.

6.1 Representation Learning: The Manifold Hypothesis

The performance of machine learning algorithms depends on the data representation. We hypothesize that this is because different representations can entangle or hide different explanatory factors of variation in the data. The features in your data are essential to the predictive models and will influence the resulting outcome. The quality and quantity of the features will have great influence on whether the model is good or not. In machine learning, feature learning or representation learning [23] is a set of techniques that allows a system to automatically discover the representations needed for feature detection or classification from raw data. This replaces manual **feature engineering** and allows a machine to both learn the features and use them to

perform a specific task.

One of the most useful ways to uncover the structure in high dimensional data is to project it down to a subspace, such as a 2-D plane, where hidden features may become visible. **Manifold learning** is based on the assumption (manifold hypothesis) that the features lie on or near a lower dimensional surface in the higher dimensional coordinate space of the data. Low dimensional structures typically arise due to constraints arising from physical laws. For instance, the laws that govern the acquisition of natural images, or the formation of speech. The first observation in favor of the manifold hypothesis is that the probability distribution over images, text strings, and sounds that occur in real life is highly concentrated. Uniform noise essentially never resembles structured inputs from these domains. Many rigorous methods have been developed to prove the manifold assumption [55],[263],[249], and many others. A reported empirical study [54] of a large number of 3×3 images represented as points in $\mathcal{R}^9$ revealed that they approximately lie on a two dimensional manifold knows as the Klein bottle.

Standard dimensionality reduction techniques, such as principal component analysis (PCA) and factor analysis (FA), work well when the data lie near a linear subspace of high dimensional space. They have substantial performance loss when the data lie near a nonlinear manifold. These problems can be reformulated as optimization problems, generalizing the projection theorem in Hilbert space [20]. As seen before in Chapter 4 the tools available in Non-Euclidean geometry are quite numerous and permit a more accurate analysis. In this chapter, we recall manifold learning techniques of interest in the space of body variation, leaving a deeper formulation for a later stage.

6.2 Human Body Manifold Learning

Representation learning takes an entirely new role in the area of human body representation. Understanding the structure of the body representation can help in a better understanding of the relations between body shapes. As seen in Chapters 2, and 4, representing the human body geometries is particularly important in biometrics and biomedical science. Although the use of anthropometric measurements makes it easy for humans to analyze body geometries, this representation is insufficient when applied to unconstrained scenarios (Chapter 5). In such situations, limiting the description to a few sample body measurements makes the geometric representation confined to trivial scenarios, and error-prone.

The human body is per se a 2D manifold in a 3D space and is a non-rigid object that can assume a variety of shapes due to body composition and pose. Moreover, for each individual, the body is also subject to change over time, due to growth or nutrition changes. The space of all body variations is high dimensional, making the design of machine learning algorithms complex, and computationally inefficient. Although, the geometric information, like natural images acquired by cameras, being governed by the laws of physics verify the manifold assumption.

6.2.1 Human Body Manifold

In the area of skinned parametric body modeling, we have seen different parametric models that permit accurate parametrization of poses as well as the shapes (See Section 5.1.1). However, one of the major problems is the availability of 3D body mesh with large variations in body shape. Previous work, leveraging the CAESAR [244] dataset, were able to learn a low dimensional manifold of the model parameters. In [96], Freifeld et al. showed how to characterize the set of all possible deformations in a human body using a Lie manifold. This approach provides an elegant solution to the representation of the space of variations for subjects from different

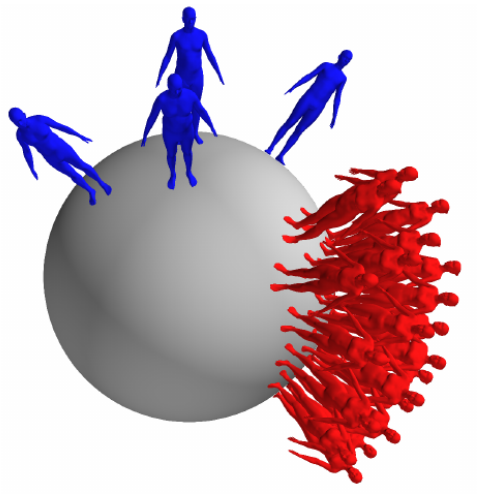


Figure 6.1: Visualization of the Body Manifold from Freifeld et.al. [97].

classes (e.g., gender, weight, etc.). In [97], they extended the previous framework using parallel transport as transfer learning method to improve the learning of datasets with missing subjects (Figure 6.1). The method is based on the Levi-Civita (LC) connection, a fundamental tool in Differential Geometry [25]. This construction leads to an ordinary differential equation (ODE) whose solution coincides with the LC parallel transport. Spectral decomposition (Chapter 4) can be used to transfer the style [100] between individuals. The results are quite impressive [32], though the method is affected by numerical instabilities.

Recently, given the importance of 3D acquisition devices like the lidar [70] devices, the interest has shifted toward raw data like the point cloud, instead of meshes. Some recent exciting works based on this idea have adopted new optimization techniques using Deep learning [230],[2]. Although these are promising ideas, they cannot be compared to traditional methods using meshes, or against the performance of Convolutional Neural Networks (CNN) on 2D data. For instance, the astonishing performance of **very deep** network on the challenging Imagenet challenge [301]).

Moreover, although there have been tremendous efforts to replace the convolution operation

in non-Euclidean spaces, some issues are still open and are far from being resolved [267]. For this reason, we believe that traditional natural image representation, with its efficiency, and the current maturity level of optimization algorithms is predominantly imposing the State-of-The-Art (SoTA) in representation learning.

Human body modeling from natural images is a more difficult task since part of the shape information is not accessible. Moreover, illumination, occlusions, and common distortions make the task harder to solve. In this area, there have been important works, such as deformable part models [90], and more recently, using CNNs [209].

We leverage the 2D rendering of the Virtual NHANES dataset to study the space of body representations. In this space, we hope to find useful relations that permits a fast semantic characterization of different body attributes (e.g., body proportions). The present work is the first of its kind, extending our understanding of human body representations, and using the human body shape manifold to address various questions on body shape semantic analysis.

6.3 A Generative Model Approach for Human Body Semantics

In this section, we introduce the generative models used to learn a latent representation of the human body. We are particularly interested in learning a structured latent space of the human body variations, able to give a disentangled representation of the attributes (weight, WHR). Given the vast literature we focus on the recent architectures: Generative Adversarial Networks (GANs), Variational Autoencoder (VAEs), leaving a more detailed description to more recent surveys [103]

6.3.1 Generative Models

In probability and statistics, a generative model is a method for generating all values of a process. Generative models are used in machine learning for either modeling data directly (e.g., modeling observations drawn from a probability density function), or as an intermediate step to forming a conditional probability density function. Generative models are typically probabilistic, specifying a joint probability distribution over observation and target (label) values. A conditional distribution can be formed from a generative model through Bayes' rule.

Generative Neural Networks (GNNs) are trained to produce samples that resemble the training set. Contrarily to Deep Neural Networks (DNNs), the number of model parameters is significantly smaller than the training data. Thus the models are forced to discover efficient data representations. These models are sampled from a set of latent variables in a high dimensional space, here called a **latent space**. Learned latent representations often also allow **semantic operations** with vector space arithmetic.

A GNN model includes an encoder to map from the feature space into a latent space, and a decoder, to map from the latent space back into the feature space. If the encoder-decoder transformation is an identity function, the goal is to reconstruct the input through the model. This network architecture, called **autoencoder** [293] (**AE**), has been at the center of research in neural networks in the last decade.

Many different variations of the original formulation have been proposed, to name a few: regularized, sparse, and contractive **AE**. The **AE** per se is not a generative model, but it can be easily modified to generate instances represented by a vector (**z**) sampled from the latent space. The simplest solution is the transformation of the deterministic latent vector in a sample drawn from a given probability distribution.

Today, two popular generative models for image data are the Variational Autoencoder (VAE [147]) and the Generative Adversarial Network (GAN [104]). VAEs can be easily interpreted as proba-

bilistic graphical models [150] with the objective of maximizing a lower bound on the likelihood of the data. GANs instead formalize the training process as a game between two adversaries: a generative network and a separate discriminative network.

Though these two frameworks are very different, both construct high dimensional latent spaces that can be sampled to generate images resembling training set data. Moreover, these latent spaces are highly structured and can enable complex operations on the generated images by simple vector space arithmetic in the latent space [158].

Variational Autoencoder (VAE)

VAEs are specified by a parametric generative model $p_\theta(x \mid z)$ of the visible variables x given the latent variables z, a prior $p(z)$ over the latent variables and an approximate inference model $q_\phi(z \mid x)$ over the latent variables given the visible variables. It can be shown that [147]:

$$\log p_\theta(x) \geq -\mathrm{KL}(q_\phi(z \mid x), p(z)) + \mathrm{E}_{q_\phi(z|x)} \log p_\theta(x \mid z). \tag{6.1}$$

where the right hand side of Eq. 6.1 is called the variational lower bound or evidence lower bound (ELBO). If there is ϕ such that $q_\phi(z \mid x) = p_\theta(z \mid x)$ we would have

$$\log p_\theta(x) = \max_\phi \{-\mathrm{KL}(q_\phi(z \mid x), p(z)) + \mathrm{E}_{q_\phi(z|x)} \log p_\theta(x \mid z)\}. \tag{6.2}$$

However, in general, this is not true, so that we only have inequality in Equation (6.2). When performing maximum-likelihood training, our goal is to optimize the marginal log-likelihood

$$\mathrm{E}_{p_\mathcal{D}(x)} \log p_\theta(x), \tag{6.3}$$

where $p_{\mathcal{D}}$ is the data distribution. Unfortunately, computing $\log p_\theta(x)$ requires marginalizing out z in $p_\theta(x, z)$ which is usually intractable. Variational Bayes uses inequality (6.1) to rephrase the intractable problem of optimizing Equation (6.3) into

$$\max_\theta \max_\phi \mathrm{E}_{p_{\mathcal{D}}(x)} \Big[-\mathrm{KL}(q_\phi(z \mid x), p(z)) + \mathrm{E}_{q_\phi(z|x)} \log p_\theta(x \mid z) \Big]. \tag{6.4}$$

Due to inequality (6.1), we still optimize a lower bound to the true maximum-likelihood objective (6.3).

The quality of this lower bound depends on the expressiveness of the inference $q_\phi(z \mid x)$. Usually, $q_\phi(z \mid x)$ is taken to be a Gaussian distribution with diagonal covariance matrix whose mean and variance vectors are parametrized by neural networks with x as input [147, 242]. While this model is very flexible in its dependence on x, its dependence on z is very restrictive, limiting the quality of the resulting model. Indeed, it was observed that applying standard Variational Autoencoders to natural images often results in blurry images [158].

Generative Adversarial Networks (GANs)

The basic idea of GANs is to set up a game between two players: the *generator* G, and the *discriminator* D (Figure 6.2). GANs are structured probabilistic models [103] with latent variables $\boldsymbol{z}$ and observed variables $\boldsymbol{x}$. The generator G creates samples that are intended to come from the same distribution as the training data. The discriminator D, instead, examines samples to determine whether they are real or fake. The discriminator is a function D that takes $\boldsymbol{x}$ as input and uses $\boldsymbol{\theta}^{(D)}$ as parameters. The generator is defined by a function G that takes $\boldsymbol{z}$ as input and uses $\boldsymbol{\theta}^{(G)}$ as parameters. Generator and discriminator are usually deep neural networks. The discriminator tries to minimize $J^{(D)}\left(\boldsymbol{\theta}^{(D)}, \boldsymbol{\theta}^{(G)}\right)$ while optimizing only $\boldsymbol{\theta}^{(D)}$. The generator, instead tries to minimize $J^{(G)}\left(\boldsymbol{\theta}^{(D)}, \boldsymbol{\theta}^{(G)}\right)$ while optimizing only $\boldsymbol{\theta}^{(G)}$. In this

scenario, each network's cost depends on the other network's parameters, but each one cannot control the other's parameters. This scenario is more intuitive to describe as a game rather than as an optimization. In game theory, the solution of a game is the Nash equilibrium [236]. A Nash equilibrium $(\boldsymbol{\theta}^{(D)}, \boldsymbol{\theta}^{(G)})$ is reached when $J^{(D)}$ falls in a local minima with respect to $\boldsymbol{\theta}^{(D)}$ and a local minimum of $J^{(G)}$ with respect to $\boldsymbol{\theta}^{(G)}$.

The discriminator learns using traditional supervised learning techniques, dividing inputs into two classes (real or fake). The generator is trained to fool the discriminator. GANs make approximations based on using supervised learning to estimate a ratio of two densities. The GAN approximation is subject to the failures of supervised learning: overfitting and underfitting. In principle, with perfect optimization and enough training data, these failures can be overcome. Other models make other approximations that have other failures.

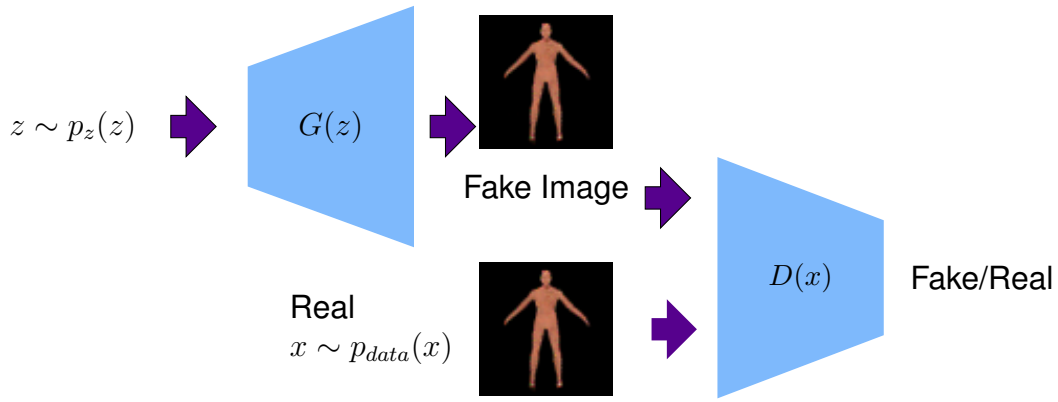


Figure 6.2: Generative Adversarial Network architecture.

The relation between the two networks, is "adversarial", given the optimization race, but also "cooperative" since the discriminator estimates this ratio of densities and then freely shares this information with the generator. From this point of view, the discriminator is more like a teacher instructing the generator in how to improve than an adversary. The discriminator strives to make $D(G(\boldsymbol{z}))$ approach zero while the generator strives to make the same quantity approach

1. If both models have sufficient capacity, then the Nash equilibrium of this game corresponds to the $G(\boldsymbol{z})$ being drawn from the same distribution as the training data, and $D(\boldsymbol{x}) = \frac{1}{2}$ for all $\boldsymbol{x}$.

A generative model must be able to generate a whole series of different outputs, for example, different faces, or different bedroom images. A set of latent variables $\mathbf{z}_i$ is drawn at random every time the model needs to generate an output. These latent variables are fed to a generator G that produces an output $\hat{x}$ (e.g., an image) $\hat{x}_i = G(\mathbf{z}_i)$. Different drawings of the latent variable result in different images being produced and the **latent variable can be seen as parameterizing the set of outputs**.

The discriminator's cost, $J^{(D)}$

The discriminator cost functions, $J^{(D)}$, used in all the GANs implementations is always the same. Instead, they vary for the cost function used for the generator, $J^{(G)}$. The cost used for the discriminator is a commonly used binary cross-entropy (BCE):

$$J^{(D)}(\boldsymbol{\theta}^{(D)}, \boldsymbol{\theta}^{(G)}) = -\frac{1}{2}\mathrm{E}_{\boldsymbol{x}\sim p_{\mathrm{data}}} \log D(\boldsymbol{x}) - \frac{1}{2}\mathrm{E}_{\boldsymbol{z}} \log\left(1 - D\left(G(z)\right)\right). \tag{6.5}$$

The difference from a common binary classifier is that is trained on two mini-batches of data; one coming from the dataset, where the label is 1 for all examples, and one coming from the generator, where the label is 0 for all examples. This training modality permits to estimate the ratio $p_{\mathrm{data}}(\boldsymbol{x})/p_{\mathrm{model}}(\boldsymbol{x})$ at every point $\boldsymbol{x}$, enabling us to compute a wide variety of divergences and their gradients. This is the main difference that sets GANs apart from variational Autoencoders and Boltzmann machines.

The generator's cost, $J^{(G)}$

For the Nash equilibrium the game is a **zero-sum game**, in which the sum of all player's costs is always zero:

$$J^{(G)} = -J^{(D)}. \tag{6.6}$$

Since $J^{(G)}$ is directly related to $J^{(D)}$, we can write the entire cost as a **value function** specifying the discriminator's payoff:

$$V\left(\boldsymbol{\theta}^{(D)}, \boldsymbol{\theta}^{(G)}\right) = -J^{(D)}\left(\boldsymbol{\theta}^{(D)}, \boldsymbol{\theta}^{(G)}\right).$$

Zero-sum games can be interpreted as **minimax** games because their solution involves minimization in an outer loop and maximization in an inner loop:

$$\boldsymbol{\theta}^{(G)*} = \arg\min_{\boldsymbol{\theta}^{(G)}} \max_{\boldsymbol{\theta}^{(D)}} V\left(\boldsymbol{\theta}^{(D)}, \boldsymbol{\theta}^{(G)}\right).$$

Interestingly, the minimax interpretation can have a deeper connection with traditional optimization approaches. In fact, it resembles minimizing the **Jensen-Shannon** divergence between the data and the model distribution, that converges to its equilibrium if both players' policies can be updated directly in function space. In practice, the players are represented with deep neural nets and updates are made in parameter space, so these results, which depend on convexity, do not apply.

6.3.2 The DCGAN Architecture

Beside the Goodfellow et al. original paper [104], adversarial architectures for vision are loosely based on the DCGAN architecture [233]. DCGAN stands for "Deep Convolutional GAN". Figure 6.2 shows the basic architecture, the generator ($G(\mathbf{z})$), and the discriminator ($D(\mathbf{x})$).

The key insights of the DCGAN architecture are:

- Use of batch normalization [132] layers in most layers of both the discriminator and the generator, with the two mini-batches for the discriminator normalized separately. The last layer of the generator and first layer of the discriminator are not batch normalized, so that the model can learn the correct mean and scale of the data distribution. See Figure 6.3.
- The overall network structure is mostly borrowed from the all-convolutional net [271]. This architecture contains neither "pooling" nor "unpooling" layers. When the generator needs to increase the spatial dimension of the representation it uses transposed convolution with a stride greater than 1.
- The use of the Adam optimizer rather than Stochastic Gradient Descent (SGD) with momentum.

DCGANs can generate high-quality images when trained on restricted domains of images, such as images of bedrooms, faces, and as we will show, bodies. DCGANs also clearly demonstrated that GANs learn to use their latent code in meaningful ways, with simple arithmetic operations in latent space, having a clear interpretation of arithmetic operations on semantic attributes of the input.

6.4 Method: Creating New Body Shapes

The GAN architecture permits us to "learn" a generator as a map from a low dimensional latent space to a high dimensional space like images, audio, and video spaces. Moreover, it has been proved [233] that the unsupervised adversarial learning can learn a "structured latent space", with different semantics and stratification. Once the generator G has been learned, we can generate new bodies $\hat{x}_i$ by just sampling new latent vectors $\mathbf{z}_i$ from the latent distribution

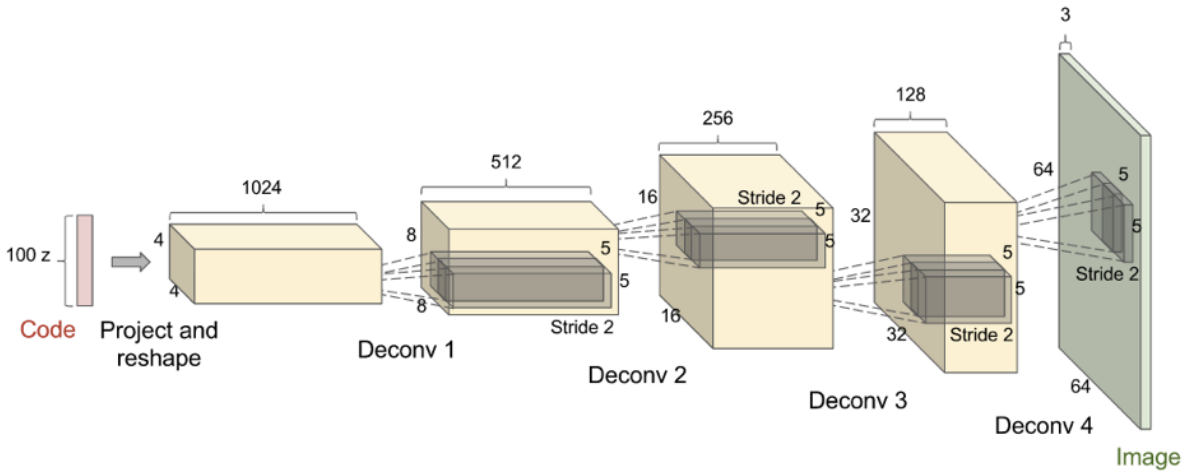


Figure 6.3: The generator network used by DCGAN. Figure reproduced from [233].

Z, and feed the generator. Our focus in this work is the study of the *sampling operation as an exploration on the manifold of the human bodies*. Given bodies with different geometries (different WHR, stature, etc.), we want to find a particular sampling operation that captures a specific relationship between the generated bodies (e.g., increasing WHR, constant WHR, etc.). Thus, the sampling is performed in a way that enforces the relationship of interest between the bodies generated by the sampling operation on the latent space.

6.4.1 Latent Space Z

We assume a latent space, generated by a normal or uniform distribution $Z \sim \{\mathcal{N}(0,1), \mathcal{U}(-1,1)\}^d$ with dimension d. On this space the sampling method will draw a batch S of latent vectors $\mathbf{z}_i \in S \subset Z, i = 1, \ldots, n$. The desired outcome is the generation of bodies $\hat{x}_i \in G(S)$ where for $\mathbf{z}_A \approx \mathbf{z}_B \rightarrow \hat{x}_A = G(\mathbf{z}_A) \approx \hat{x}_B = G(\mathbf{z}_B)$. We use the WHR as measure of body similarity, thus for $\mathbf{z}_A \lesseqgtr \mathbf{z}_B \rightarrow WHR(\hat{x}_A) \lesseqgtr WHR(\hat{x}_B)$. This formulation permits to compare the generated bodies efficiently. Traditional works on faces, or scene generation using GANs, instead, rely on the visual inspection by the human. The presented framework can be seen as a

supervised system, contrarily to GANs framework that are mostly unsupervised.

Both VAEs, and GANs operate by placing a prior distribution over a latent space $p(Z)$ and learning a mapping from the latent space, Z, to the space of the observed data $\hat{x} \in X$. Thus the latent space will have some areas dense with bodies. Unfortunately, given the unsupervised nature of these generative models, it is difficult to understand the structure of the latent space. However, we can expect to find areas with low prior, creating some holes in the manifold. We also expect different behaviors of the generator given the non-Euclidean nature of the latent space.

Sampling Techniques

Generative models are often evaluated by examining samples from the latent space. Frequently used techniques are **random sampling** and **linear interpolation**. These can result in sampling the latent space from locations very far outside the manifold of probable locations. When sampling the latent space is preferable to be close to locations that are more likely given the prior of the model. This technique has been used in the original VAE method [147] which adjusted sampling through the inverse CDF of the Gaussian to accommodate the Gaussian prior. The second principle is to consider that the dimensionality of the latent space is often artificially high and may contain dead zones that are not on the manifold learned during training [191]. In this work we focus on two interpolation techniques: a **linear interpolation**, and a **spherical linear interpolation** (See Figures 6.4, 6.5), in addition to the traditional **random sampling** technique.

Interpolation

Interpolation is used to find new points between two known locations in latent space (Figure 6.4). It has been used as a way of demonstrating that a generative model is not directly

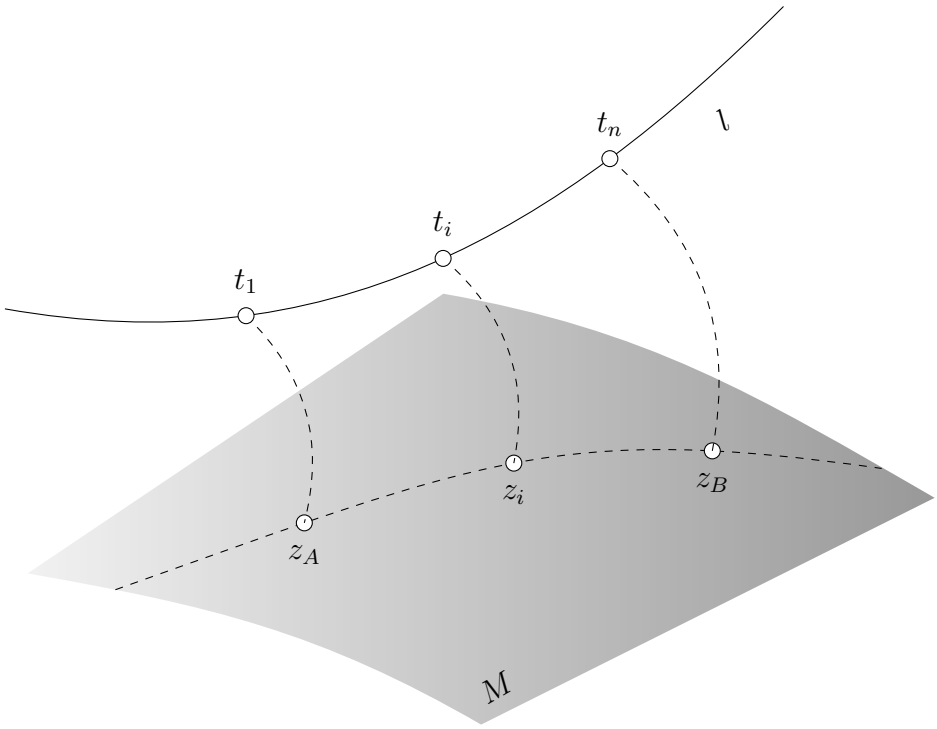


Figure 6.4: Sampling operation on the manifold.

memorizing the training examples, but is learning the manifold representation of the data [233].

Linear Interpolation Given two samples $\mathbf{z}_A, \mathbf{z}_B \in Z^d$ as the extremes of the line, the intermediate samples are computed as $\mathbf{z}_i = \mathbf{z}_B \cdot t_i + \mathbf{z}_A \cdot (1 - t_i)$ with $t_i = 0, \ldots, n$, where n is the batch size. Linear interpolation, (see Figure 6.4) is easily understood and implemented, but often inappropriate as the latent spaces of most generative models are high dimensional (> 50 dimensions). In such a space, linear interpolation traverses locations that are extremely unlikely given the prior. Let's consider a 100 dimensional space with the Gaussian prior $\mathcal{N}(0, 1)$. All random vectors will have a length very close to 10 (standard deviation < 1). However, linearly interpolating between any two samples will usually result in the vector magnitude decreasing

from roughly 10 to 7 at the midpoint, which is over 4 standard deviations away from the expected length.

Spherical Linear Interpolation Considering the latent space as a 2D manifold (e.g., sphere surface, Figure 6.5) we can consider the great circle on the sphere surface. We can sample points on the surface of the sphere on the path of the great circle. In particular, using *slerp* **s**pherical **l**inear int**erp**olation [265] allows us to move at constant-speed along a unit-radius great circle arc, given the ends and an interpolation parameter between 0 and 1. This formula is a symmetric weighted sum, thus any point on the curve must be a linear combination of the endpoints.

$$Slerp(\mathbf{z}_A, \mathbf{z}_B; t) = \frac{\sin([1-t]\Omega)}{\sin(\Omega)} \cdot \mathbf{z}_A + \frac{\sin([t\Omega])}{\sin(\Omega)} \cdot \mathbf{z}_B \tag{6.7}$$

where $\mathbf{z}_A, \mathbf{z}_B$ are the endpoints of the arc, and t is the parameter, $0 \leq t \leq 1$. We compute Ω as the angle subtended by the arc so that $\cos(\Omega) = \mathbf{z}_A - \mathbf{z}_B$, the n-dimensional dot product of the unit vectors from the origin to the ends. A *slerp* path is an equivalent in spherical geometry of a path along a line segment in the plane. Thus a great circle is a spherical geodesic [265].

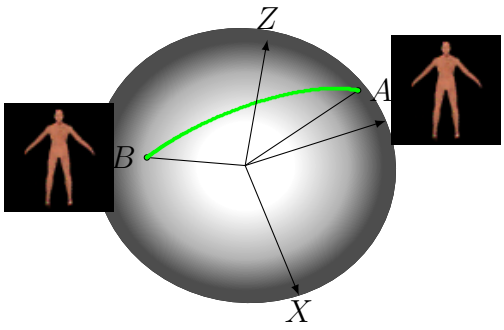


Figure 6.5: Sperical Interpolation of two samples $\mathbf{z}_A, \mathbf{z}_B$ on the latent space.

Random Sampling Random sampling is interesting for the simplicity of the method, but also because is a simple test that can tell if the latent space is "biased" toward some of the modes. It's also interesting because it can shows bodies not accessible with other techniques.

6.4.2 Evaluation Network

The evaluation of generative models is a challenging task. The fundamental difficulty resides in the ill formulation of this task. The generation of new images, never seen before, but generated from the same distribution, can be cast as density estimation. For density estimation, log-likelihood (or equivalently, the KL divergence) has been the standard for training and evaluating generative models. However, the likelihood of many exciting models is computationally intractable. Generative models are also often compared regarding properties more readily accessible than likelihood. For instance, visualizations of model samples, interpretations of model parameters, Parzen window estimates of the models log-likelihood, and evaluations of model performance in surrogate tasks such as denoising or missing value imputation.

Given the objective of the generator: create new bodies with realistic shapes, we believe that a reliable evaluation method would be to measure some critical body semantics, together with the visual inspection of the results. We use the Waist-to-Height ratio (WHR) as a body shape indicator. This ratio is directly connected with the body appearance (Chapter 5), and do not suffer from the drawbacks occurring with BMI, and BSA. To estimate this indicator, we train a convolutional neural network able to regress this value from the natural images. However, training the network on the same dataset used to learn the generative model will most likely overfit the regressor. Thus, we decided to follow a **transfer learning** approach.

WHR Regressor Network

We use a pre-trained model for face classification on the CelebA dataset [181], composed of 4 convolutional layers (4×4 kernels size, 2×2 stride, and leaky-Relu activation function, and batch normalization). The use of the pre-trained model is motivated by the transfer learning assumption in vision. Neural networks mimicking the human vision system, "store" in the lower layers the basic components of images, such as corners and edges. Higher layers, instead, store high-level knowledge about the image. Transfer learning assumes that the basic structures in vision are common for the natural images, thus it is correct to train on larger dataset (greater generalization), but we test on a different one. Before fine-tuning on the WHR regression task, we modify the network adding a convolutional layer with leaky ReLU [300], another fully connected layer with ReLU activation function, and a Sigmoid function. Given the image size (64x64), and the number of training data, we decided not to use a very deep neural network. To avoid memorizing effects of the network we freeze the pre-trained layers, and we train the remaining layers. For these layers, a weight initialization with normal statistic is used. The training is conducted using the Adam algorithm [148] with the mean squared loss function. We use a schedule policy for the learning rate and decay rate. Figure 6.6 shows the regressor network architecture used for the evaluation task.

6.5 Results

The body generator is based on the DCGAN [233] architecture, while the regressor is a seven layer network. The experiments consist in verifying that certain characteristics of the generated bodies characteristics are consistent with some expected patterns. Generative models can produce a latent space that is not **tightly packed**, and the dimensionality of the latent space is often set artificially high. As a result, the manifold of trained examples can be a subset of the

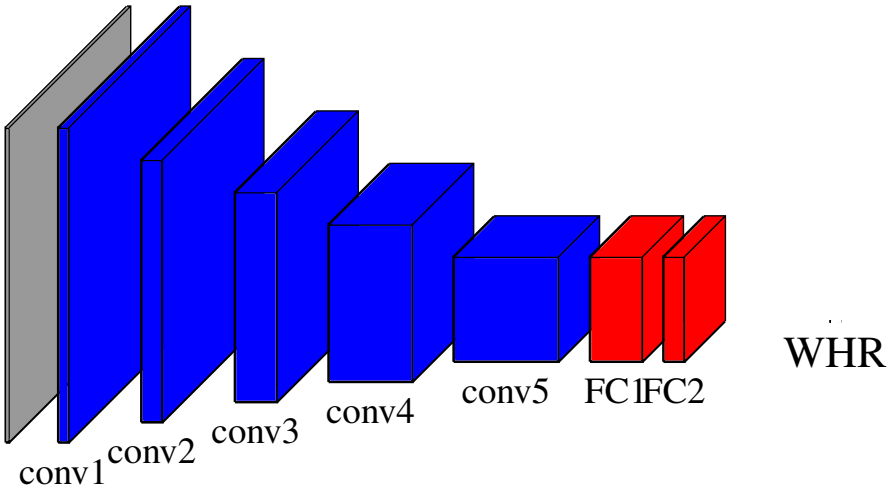


Figure 6.6: WHR regressor network.

latent space after training, resulting in dead zones, or abrupt changes in the expected prior. This situation can be easily verified by observing Figure 6.7. From a body with decreasing WHR, traveling on the great circle, the prior changes, generating child bodies with high WHR (around subject 58). A similar situation can be observed for the linear interpolation in Figure 6.10. We sample batches of vectors in the latent space with a given prior (normal or uniform distributions). We feed the generator network with these batches, and we evaluate the WHR of the corresponding set of images. When close points on the latent space generate bodies with close WHR values, we can say that those points lie on a high prior area.

Spherical Linear Interpolation In Figure 6.7 we report the generated bodies relative to samples $\mathbf{z}_i$ obtained with a spherical interpolation (slerp) on the latent space $Z \sim \mathcal{N}(0, 1)$. Interestingly the bodies relative to these latent vectors have decreasing WHR values, as we can see from Figure 6.7 (top) until the interpolation gets close to the "manifold edge". When sampling in this area with low prior, the generator gets unstable, generating noisy, shaded bodies (Fig-

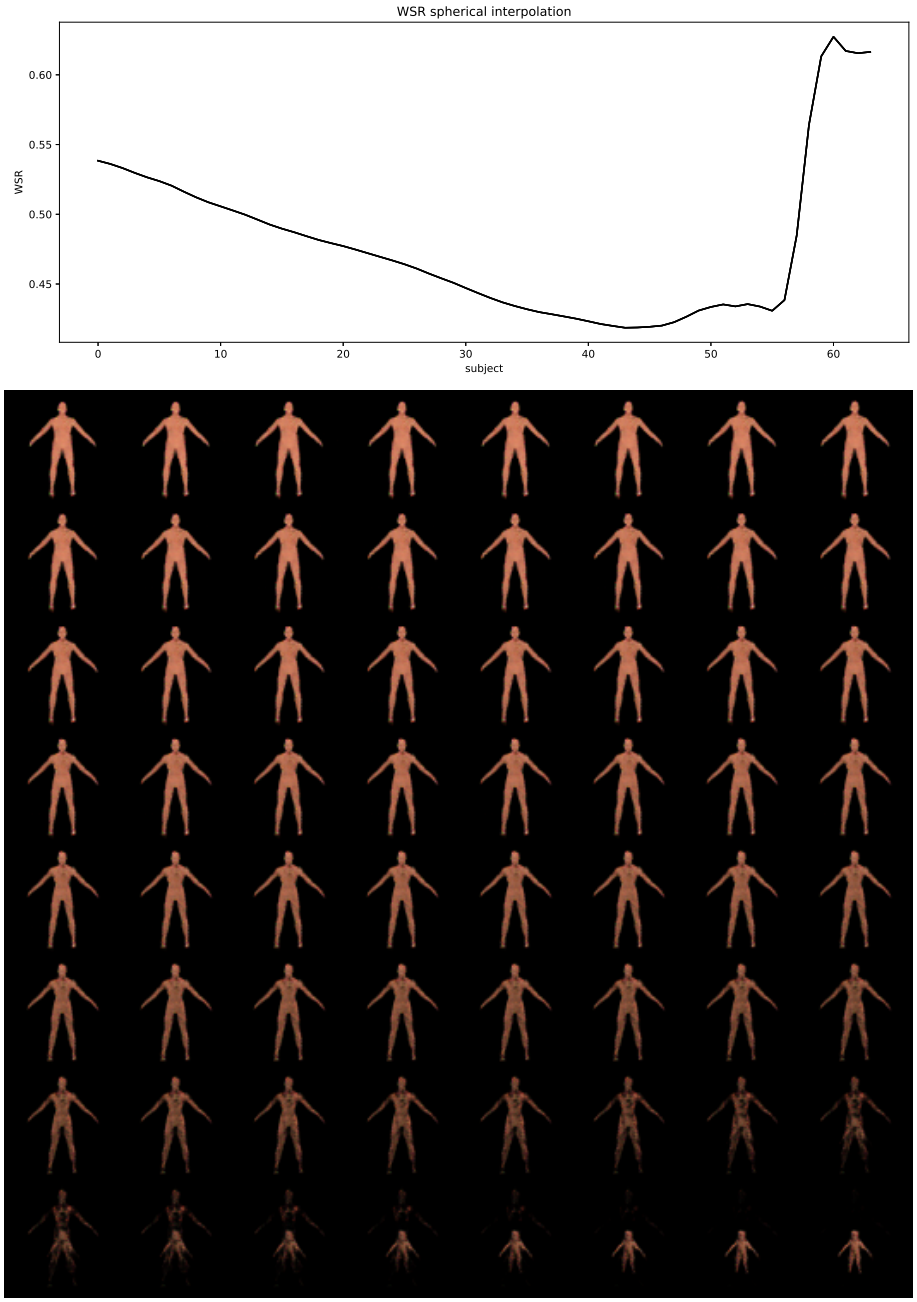
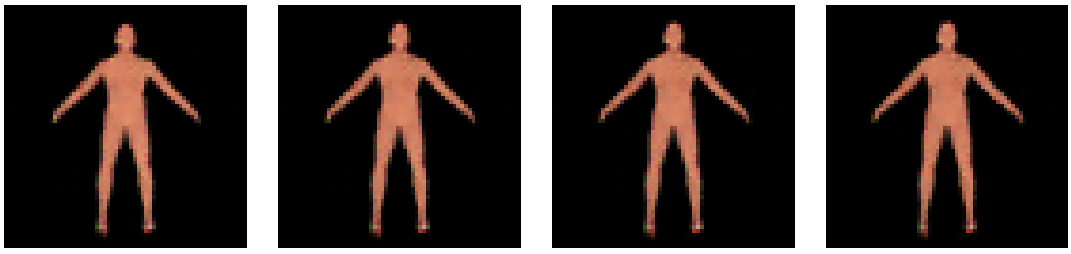


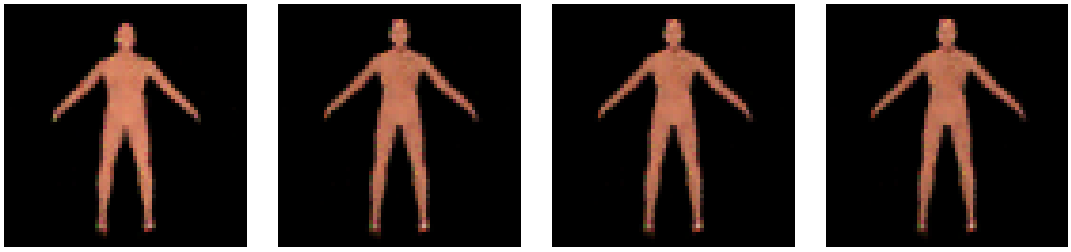
Figure 6.7: Results using Spherical Linear Interpolation on the Latent Space. Above the WHR of the batch of subjects. Below the generated images. Images are labeled row wise: from left to right, and top to bottom.

ure 6.9). This behavior can be explained by different hypothesis. First, the used dataset has a low number of small bodies, creating a bias toward adult bodies that constitute the main modes, thus associated with the areas of high priors. Second, for small skinny subjects the network cannot generate clear bodies. The used CNN has convolutional filters with large receptive fields and hence lead to coarse outputs when generates pixel-level objects. Although the absence of max-pooling layers in the network model, the current implementation is not be able to extract fine-grained structures in the image, like face, and muscle. Last, but not least, we notice that there can be an entanglement problem between skinny bodies and the shading augmentation present in the Virtual NHANES rendered dataset. Unfortunately, this last drawback is more difficult to prove since we need more exploratory processing of the training and generated images. Even with these drawbacks, we can conclude that the spherical interpolation permits to generate bodies with varying WHR, similar to an increase of body weight. This behavior is similar to a variation of body fat, thus movement on the x-axis in Figures 4.7-4.8.

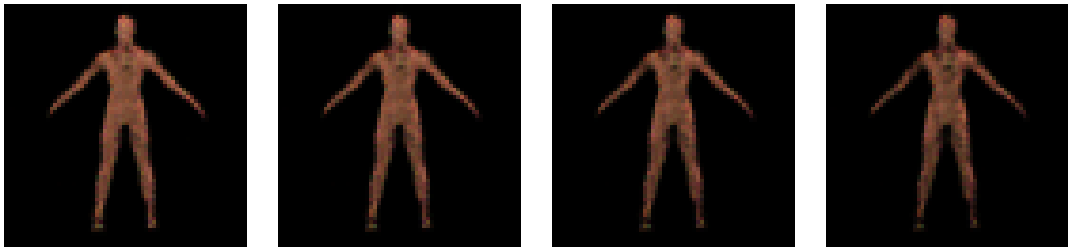
Linear Interpolation Using a linear interpolation in nonlinear high dimensional space can lead to suboptimal results since it has a high probability to fall in areas with low prior. However, when the extremes are close enough to be contained in a small area, we can find interesting results. In Figure 6.10 we can see, after some samples outside the manifold, that the generated bodies have almost constant WHR. Observing the generated bodies, we can see some relevant differences: subjects in 3rd to 5th row in Figure 6.10 are closer to a male shape (hip circumference is small). Instead, 6th to 8th rows are closer to a female shape. In general, while the spherical interpolation walk on the manifold of similar subjects with varying WHR, the linear interpolation crosses between male and female subjects with constant WHR. Therefore, in a sense, the spherical interpolation answers the question of walking in the manifold from point A to point B, along the path having an increasing or decreasing value (depending on the seman-



(a) 0.538 (b) 0.526 (c) 0.520 (d) 0.516

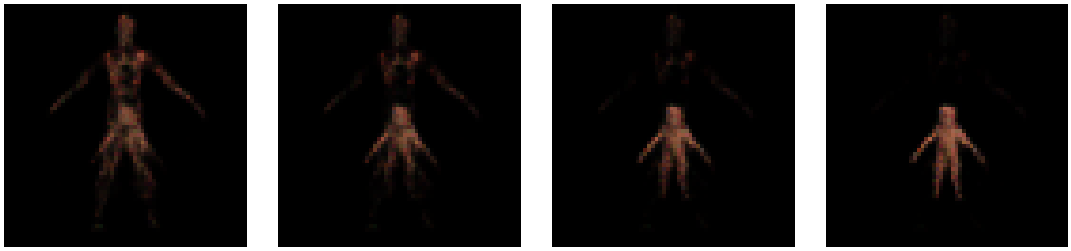


(e) 0.538 (f) 0.484 (g) 0.481 (h) 0.438



(i) 0.419 (j) 0.418 (k) 0.420 (l) 0.426

Figure 6.8: Some results of Spherical Linear Interpolation with relative WHR.



(a) 0.438 (b) 0.563 (c) 0.627 (d) 0.615

Figure 6.9: Spherical interpolation: Examples of bodies outside the high prior manifold. Numbers indicate the WHR values.

tic relation of interest). The linear interpolation, on the other hand, addresses the question of walking on a path with a constant value for the relation of interest.

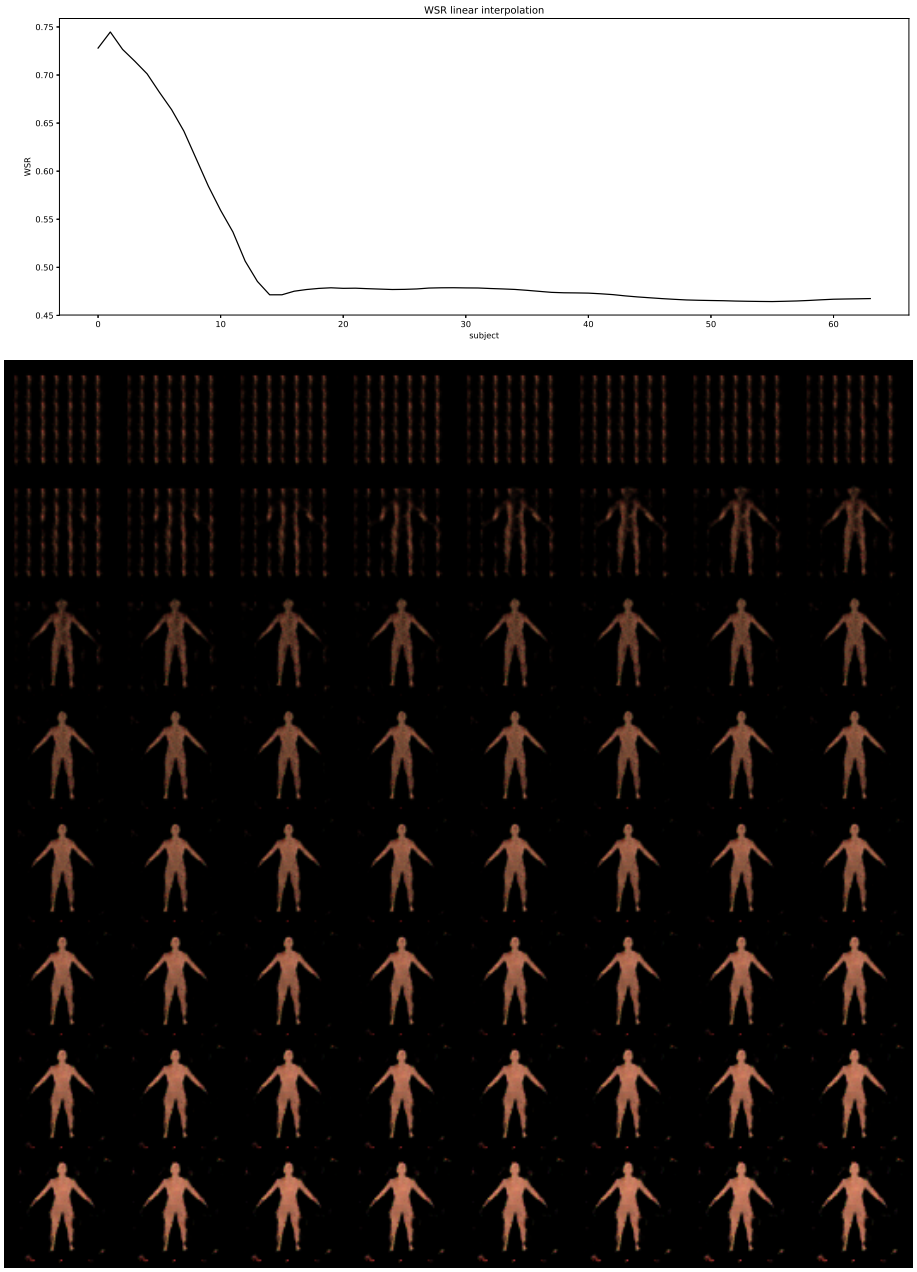


Figure 6.10: Linear Interpolation on the Latent Space. Above the WHR of the batch of subjects, below the generated images. Images are labeled row-wise: from left to right, and top to bottom.

Random Sampling For some applications, we may need to create a random population of subjects from a set of random vectors. In this situation, we can use random sampling on the

latent space. However, given the implicit structure of the latent space, with small areas with high priors, we can easily overshoot these locations, generating an unrealistic population. In Figure 6.11(top) we can see the WHR distribution of a random batch of bodies. Although there are a few cases with low prior and high shading, the majority of the subjects span a large number of shapes and dimensions, not seen before with the interpolation experiments. This experiment proves that the generator can generate many modes (body shapes), although the manifold structure is more sparse with many locations with very high priors.

6.6 Conclusion

In this Chapter, we proposed an innovative generative model for the exploration of the human body manifold. The task is particularly challenging given the instability of the GAN architecture. To overcome the problem related to the evaluation of generative methods we designed a regressor network able to retrieve the Waist-to-Height Ratio given the generated body image. We analyzed the generator latent space as a manifold, adopting different sampling techniques. The generative method is particularly interesting because permits the inference of new bodies with a simple arithmetic operation on the latent vectors and the fast-forward pass in the generator network. We discover exciting patterns on the latent space, but we also verified the presence of low prior areas, leading to the conclusion that the learned manifold is not compact, or it can still be reduced to a lower dimensional space.

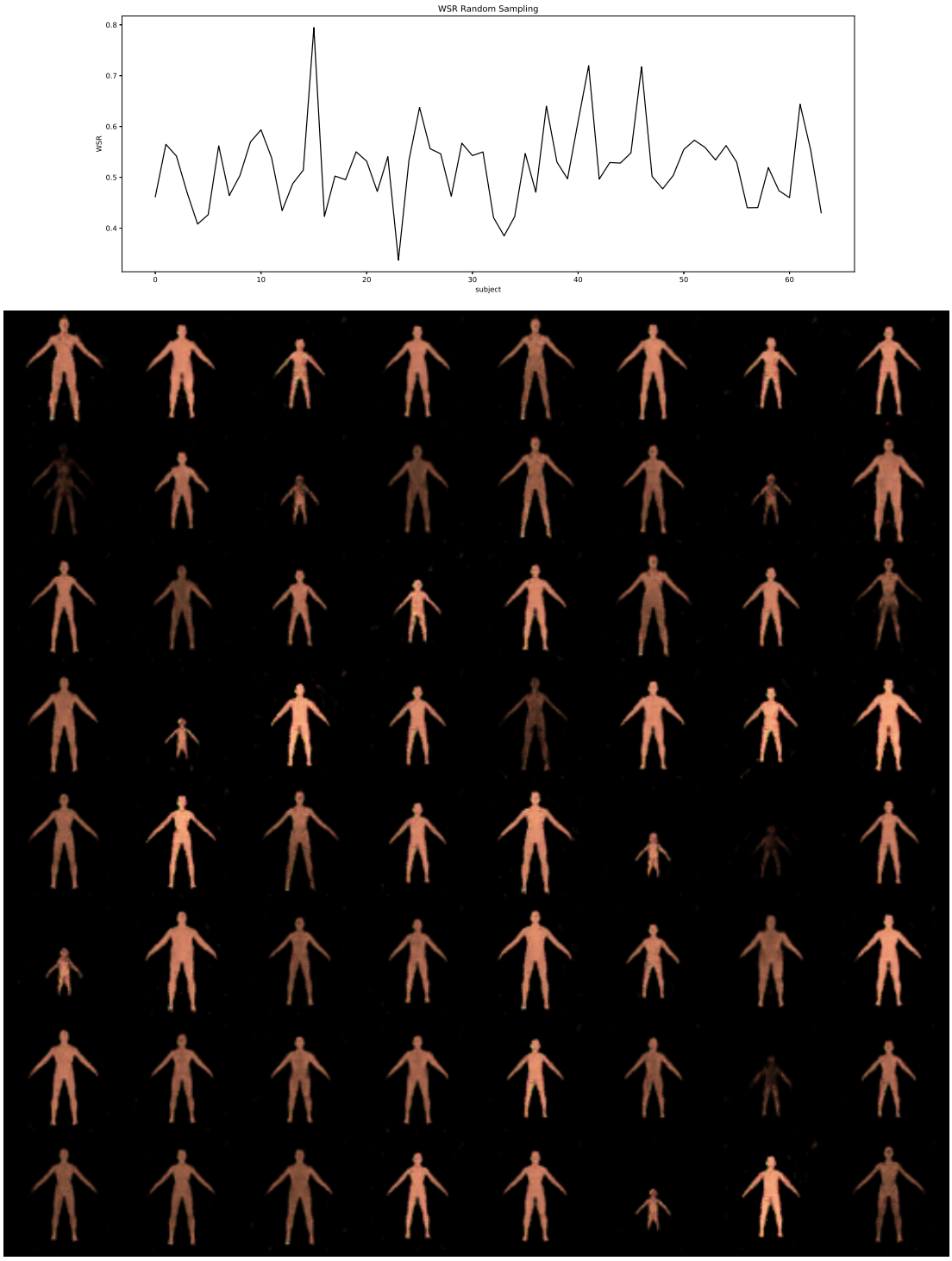


Figure 6.11: Random Sampling with Gaussian noise on the Latent Space. Images are labeled row-wise: from left to right, and top to bottom.

Chapter 7

Conclusion and Future work

7.1 Conclusion

In this thesis, we analyzed the human body variations and representation from a machine learning perspective. The human body visualization and representation has a long history dating back many centuries ago. Today we are witnessing the Artificial Intelligence (AI), and Machine Learning (ML) revolution. For modern AI/ML systems, it is important to understand the real world, and the living beings that populate it. We motivate the work focusing our attention on biometrics, and biomedical science, however, the developed techniques can cover a larger spectrum of applications.

In this work, we pay particular attention in the design of the key components of a modern ML/AI system, with many contributions in different areas. Fundamental for a data-driven system, we introduce a new method (**VirtualBODY**) able to generate a population of 3D human models with rich semantics, and detailed anthropometric measurements (Chapter 2). A unique feature of this method is the generation of body measurements, showing that the generated population is statistically comparable with populations of real subjects. This method allowed us to

generate three datasets able to tackle most common problems in human body shape analysis: high number of subjects, high variance of the dimensions (stature, weight, gender, race, etc.), and measurements under pose variation (Chapter 5).

Taking advantage of the newly introduced dataset, we proved that with a single-view RGB-D camera it is possible to infer the whole body surface area of a human. In this work, we realized a virtual environment able to simulate the physician office and the acquisition process by an RGB-D camera. The WBSA, as proved in a subsequent chapter, characterize the spectrum of the shape, and has important applications in medicine, and can also be a useful soft biometric feature.

We introduced a **Spectral Geometry** approach for body fat analysis (Chapter 4). Spectral geometry has been used before mostly for shape retrieval, but it has never been applied to body shape analysis in medical science, or in soft biometrics. 3D **spectral analysis** is based on the Laplace-Beltrami Operator (LBO). LBO has the important property of being invariant to deformations of the shape that maintain the metric on the surface (**isometry** and **quasi-isometry**). Classical body deformations due to different poses are parts of these transformations, thus making our system largely independent of the pose. This innovation permits some interesting analysis for automatic health assessment. We present a spectral method for semantic classification of Body Fat percentage (BFP). In the same chapter, we introduced some theoretical results exploring the interaction between spectral analysis, BFP, and body surface area.

A capital task in computer vision and ML is to obtain a representation invariant to some nuisances. In human body analysis, the body pose is the most critical factor that introduces uncertainty. In analysis of human anthropometry, the pose invariance is rarely analyzed. Using the **VirtualBody** framework, we conduct a detailed statistical analysis to show which body measurements can be considered invariant to a common set of pose changes (Chapter 5). This study is the first in the literature that focus the attention to pose invariance in human body shape

analysis. It can also be used for future studies with real data to improve the robustness of ML systems.

Human body variations are due not only to pose but also to growth. A common problem in medical science is to track, but also predict the variations of the human body due to some event (e.g., changes in nutrition, diseases, etc.). Traditional techniques have been based on BMI, and WBSA indices, however, are inadequate, as discussed in (Chapter 1). In Section 4.5.3 we presented an initial solution to the problem of representing a family of bodies with related or common characteristics. However, learning a CRF model can be a daunting task, given the bidirectional relation in an undirected graph. In Chapter 6, we propose a solution based on Deep Generative models. This method permits us to explore the body manifold with a simple sampling operation on the latent space. Given the fast-forward speed of the convolutional neural networks (CNN) for inference problems, we can analyze a large number of bodies in relatively short time. Moreover, the generator latent space can be analyzed with traditional vector calculus tools, as well as more complex statistical learning methods. We have reported some main results based on different sampling operations on the latent space.

7.2 Future Work

Given the exponential progress in ML and AI, and the growing interest in human-centric applications, there can be many future directions from this study. We can divide the possible future contributions in two main areas: Spectral Geometry/3D based geometric processing, and 2D computer vision techniques.

7.2.1 Spectral Geometry/3D based Geometric Processing

Spectral Geometry, based on the Laplace Beltrami operator (LBO) is still constrained due to the computational requirements for large meshes, and the inability to perform the convolution operation efficiently in 3D space. There are some new directions, based on deep learning that permit us to lower the computational burden due to the optimization process. See for instance [39]. An alternative work can be the design of a Laplace operator as a neural network layer. This basic operation is at the base of many spectral methods on graphs. Used in a 3D neural network architecture, will permit the fast extraction of essential features, that can be easily fused with traditional 2D features.

7.2.2 2D Computer Vision

Although the 3D based processing is growing at a good pace, the growth is not as fast as the 2D methods. In this work, we cited multiple times the causes of this gap. Expanding the intuition in Chapter 6, an unusual direction would be the use of more powerful techniques to analyze the generator latent space. This highly structured space contains useful information about the relationship between bodies. The proposed sampling techniques, although simple and efficient limit the possible explorations. Using more powerful statistical learning methods, we can obtain a better understanding of this space. For instance, we can design a body shape classifier based on the latent space, as well as impose different relationships between the bodies with a conditional distribution.

Another problem worth to mention is the application of these methods on the real domain. With the knowledge acquired from some recent works [202], [201] we believe we can develop new techniques to transfer the learned knowledge on synthetic data to real scenarios. This task, known as “Domain Adaptation” is particularly appealing, attracting significant attention in the

latest major computer vision, and ML conferences [216].

Bibliography

[1] Martín Abadi and et. al. Tensorflow: A system for large-scale machine learning. In *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, pages 265–283, GA, 2016. USENIX Association.

[2] P. Achlioptas, O. Diamanti, I. Mitliagkas, and L. Guibas. Representation learning and adversarial generation of 3d point clouds. *ArXiv e-prints*, July 2017.

[3] D Adjeroh, Deng Cao, M Piccirilli, and A Ross. Predictability and correlation in human metrology. In *Information Forensics and Security (WIFS), 2010 IEEE International Workshop on*, pages 1–6, 2010.

[4] A C Adler, B H Nathanson, K Raghunathan, and W T McGee. Misleading indexed hemodynamic parameters: The clinical importance of discordant BMI and BSA at extremes of weight. *Crit Care*, 16:471, 2012.

[5] P Agarwal and S Sahu. Determination of hand and palm area as a ratio of body surface area in Indian population. *Indian J Plast Surg*, 43(1):49–53, jan 2010.

[6] Brett Allen, Brian Curless, and Zoran Popović. The space of human body shapes: Reconstruction and parameterization from range scans. *ACM Trans. Graph.*, 22(3):587–594, July 2003.

[7] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. SCAPE: Shape completion and animation of people. *ACM Trans. Graph.*, 24(3):408–416, July 2005.

[8] Wolfgang Arendt and Wolfgang P. Schleich, editors. *Mathematical Analysis of Evolution, Information, and Complexity*. Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, Germany, mar 2009.

[9] M Ashwell, P Gunn, and S Gibson. Waist-to-height ratio is a better screening tool than waist circumference and BMI for adult cardiometabolic risk factors: Systematic review and meta-analysis. *Obesity Reviews : An Official Journal of the International Association for the Study of Obesity*, 13(3):275–86, March 2012.

[10] L. Assassi, M. Becker, and N. Magnenat-Thalmann. Dynamics skin deformation based on biomechanical modeling. In *25th Annual Conference on Computer Animation and Social Agents (CASA 2012),*, May 2012.

[11] Mathieu Aubry, Ulrich Schlickewei, and Daniel Cremers. The wave kernel signature: A quantum mechanical approach to shape analysis. *Proceedings of the IEEE International Conference on Computer Vision*, 2:1626–1633, 2011.

[12] Francis Bach. Sparse modeling for image and vision processing. *Foundations and Trends® in Computer Graphics and Vision*, 8(2):85–283, 2014.

[13] Sharyn D. Baker, Jaap Verweij, Eric K. Rowinsky, Ross C. Donehower, Jan H. M. Schellens, Louise B. Grochow, and Alex Sparreboom. Role of body surface area in dosing of investigational anticancer agents in adults, 1991–2001. *Journal of the National Cancer Institute*, 94(24):1883–1888, 2002.

[14] Sachchidananda Banerjee and Ashim Kumar Bhattacharya. Determination of body surface area in Indian Hindu children. *Journal of Applied Physiology*, 16(6):969–970, 1961.

[15] Dale J. Barr, Roger Levy, Christoph Scheepers, and Harry J. Tily. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68(3):255–278, April 2013.

[16] Manuel Bastioni, Simone Re, and Shakti Misra. Ideas and methods for modeling 3d human figures: The principal algorithms used by makehuman and their implementation in a new approach to parametric modeling. In *Proceedings of the 1st Bangalore Annual Compute Conference*, COMPUTE '08, pages 10:1—-10:6, New York, NY, USA, 2008. ACM.

[17] G. Baudat and F. Anouar. Kernel-based methods and function approximation. In *Neural Networks, 2001. Proceedings. IJCNN '01*, pages 1244–1249, 2001.

[18] Martin Bauer, Martins Bruveris, and Peter W. Michor. Overview of the geometries of shape spaces and diffeomorphism groups. *Journal of Mathematical Imaging and Vision*, 50(1-2):60–97, jan 2014.

[19] Regis Behmo, Nikos Paragios, and Veronique Prinet. Graph commute times for image representation. *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008.

[20] M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396, June 2003.

[21] S Belongie, J Malik, and J Puzicha. Shape context: A new descriptor for shape matching and object recognition. *NIPS*, 546:831–837, 2000.

[22] C. BenAbdelkader and L. Davis. Estimation of anthropomeasures from a single calibrated camera. In *7th International Conference on Automatic Face and Gesture Recognition (FGR06)*, pages 499–504. IEEE, 2006.

[23] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(8):1798–1828, August 2013.

[24] Jon Louis Bentley. Multidimensional binary search trees used for associative searching. *Communications of the ACM*, 18(9):509–517, sep 1975.

[25] M. Berger. *A Panoramic View of Riemannian Geometry*. Springer Berlin Heidelberg, 2007.

[26] Marcel Berger, Paul Gauduchon, and Edmond Mazet. *Le Spectre d'une Variété Riemannienne*, volume 194 of *Lecture Notes in Mathematics*. Springer Berlin Heidelberg, Berlin, Heidelberg, 1971.

[27] S. Biasotti, S. Marini, M. Mortara, and G. Patane. An overview on properties and efficacy of topological skeletons in shape modeling. In *2003 Shape Modeling International.*, pages 245–254. IEEE Comput. Soc, 2003.

[28] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., aug 2006.

[29] Henry Blackburn and David Jacobs. Commentary: Origins and evolution of body mass index (BMI): Continuing saga. *International Journal of Epidemiology*, 43(3):665–9, jun 2014.

[30] Federica Bogo, Michael J Black, Matthew Loper, and Javier Romero. Detailed full-body reconstructions of moving people from monocular RGB-D sequences. *ICCV*, 2015.

[31] Federica Bogo, Javier Romero, Matthew Loper, and Michael J. Black. FAUST: Dataset and evaluation for 3D mesh registration. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3794–3801, 2014.

[32] Davide Boscaini, Davide Eynard, Drosos Kourounis, and Michael M. Bronstein. Shape-from-Operator: Recovering shapes from intrinsic operators. *Computer Graphics Forum*, 34(2):265–274, 2015.

[33] Bernhard E. Boser, Isabelle M. Guyon, and Vladimir N. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory - COLT '92*, pages 144–152, New York, New York, USA, jul 1992. ACM Press.

[34] Leon Bottou. Stochastic Gradient Descent Tricks. *Neural Networks: Tricks of the Trade*, 1(1):421–436, 2012.

[35] Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Describing people: A poselet-based approach to attribute classification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1543–1550, 2011.

[36] A M Bronstein, M M Bronstein, and R Kimmel. *Numerical Geometry of Non-rigid Shapes*. Monographs in Computer Science. Springer New York, 2008.

[37] A M Bronstein, Michael Bronstein, B Bustos, U Castellani, M Crisani, B Falcidieno, L J Guibas, I Sipiran, I Kokkinos, V Murino, M Ovsjanikov, G Patané, M Spagnuolo, and J Sun. SHREC 2010: Robust feature detection and description benchmark. *Proc. EUROGRAPHICS Workshop on 3D Object Retrieval (3DOR)*, 2010.

[38] AM Bronstein and MM Bronstein. Shape Google: Geometric words and expressions for invariant shape retrieval. *ACM Transactions on Graphics (TOG)*, 2011.

[39] M. M. Bronstein, J. Bruna, Y. LeCun, A. Szlam, and P. Vandergheynst. Geometric deep learning: Going beyond Euclidean data. *IEEE Signal Processing Magazine*, 34(4):18–42, July 2017.

[40] Michael M. Bronstein and Iasonas Kokkinos. Scale-invariant heat kernel signatures for non-rigid shape recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1704–1711, 2010.

[41] Lucy M Browning, Shiun Dong Hsieh, and Margaret Ashwell. A systematic review of waist-to-height ratio as a screening tool for the prediction of cardiovascular disease and diabetes: 0·5 could be a suitable global boundary value. *Nutrition Research Reviews*, 23(2):247–69, Dec 2010.

[42] Joan Bruna, Wojciech Zaremba, Arthur Szlam, and Yann LeCun. Spectral Networks and Locally Connected Networks on Graphs. *Iclr*, page 14, 2014.

[43] Richard V Burkhauser and John Cawley. Beyond BMI: The value of more accurate measures of fatness and obesity in social science research. *Journal of Health Economics*, 27(2):519–29, mar 2008.

[44] Koen Buys, Cedric Cagniart, Anatoly Baksheev, Tinne De Laet, Joris De Schutter, and Caroline Pantofaru. An adaptable system for RGB-D based human body detection and pose estimation. *Journal of Visual Communication and Image Representation*, 25(1):39–52, 2014.

[45] Koen Buys, Jonas Hauquier, Cedric Cagniart, Tinne Tuytelaars, and Joris De Schutter.

Virtual data generation based on a human model for machine learning applications. In *In Proceedings of the International Digital Human modeling Conference*, 2013.

[46] Ru O Blan and Michael J Black. The naked truth: Estimating body shape under clothing. In *In Proceedings at ECCV 2008*, 2008.

[47] M. Piccardi C. Madden. Height measurement as a session-based biometric for people matching across disjoint camera views, 2005.

[48] Yaiza Canzani. Analysis on manifolds via the Laplacian. Technical report, Harvard, 2013.

[49] Deng Cao, Cunjian Chen, Donald Adjeroh, and Arun Ross. Predicting gender and weight from human metrology using a copula model. *2012 IEEE 5th International Conference on Biometrics: Theory, Applications and Systems, BTAS 2012*, pages 162–169, 2012.

[50] Deng Cao, Cunjian Chen, Marco Piccirilli, Donald Adjeroh, Thirimachos Bourlai, and Arun Ross. Can facial metrology predict gender? In *2011 International Joint Conference on Biometrics, IJCB 2011*, 2011.

[51] Steve Capell, Matthew Burkhart, Brian Curless, Tom Duchamp, and Zoran Popović. Physically based rigging for deformable characters. *Graphical Models*, 69(1):71–87, 2007.

[52] Steve Capell, Seth Green, Brian Curless, Tom Duchamp, and Zoran Popović. Interactive skeleton-driven dynamic deformations. *ACM Transactions on Graphics*, 21(3):1–8, 2002.

[53] Peter Carbonetto, Gyuri Dorkó, Cordelia Schmid, Hendrik Kück, and Nando De Freitas.

Learning to recognize objects with little supervision. *International Journal of Computer Vision*, 77(1-3):219–237, 2008.

[54] Gunnar Carlsson. Topology and data. *Bulletin of the American Mathematical Society*, 46(2):255–308, 2009.

[55] L. Cayton. Algorithms for manifold learning. *Univ. of California at San Diego Tech. Rep*, 2005.

[56] Centers for Disease Control and Prevention (CDC). National Center for Health Statistics (NCHS). National Health and Nutrition Examination Survey Data., 1999-2013.

[57] Cha Zhang and Tsuhan Chen. Efficient feature extraction for 2D/3D objects in mesh representation. In *Proceedings 2001 International Conference on Image Processing*, volume 2, pages 935–938. IEEE, 2001.

[58] John E Chadwick, David R Haumann, and Richard E Parent. Layered construction for deformable animated characters. *Computer Graphics*, 23(3), 1989.

[59] Sharat Chandran. Introduction to kd-trees. *University of Maryland Department of Computer Science*, 2004.

[60] Ding-Yun Chen, Xiao-Pei Tian, Yu-Te Shen, and Ming Ouhyoung. On Visual Similarity Based 3D Model Retrieval. *Computer Graphics Forum*, 22(3):223–232, sep 2003.

[61] Yinpeng Chen, Zicheng Liu, and Zhengyou Zhang. Tensor-based human body modeling. In *Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '13, pages 105–112, Washington, DC, USA, 2013. IEEE Computer Society.

[62] M. M. Cheng, V. A. Prisacariu, S. Zheng, P. H. S. Torr, and C. Rother. Densecut: Densely connected crfs for realtime GrabCut. *Computer Graphics Forum*, 34(7):193–201, Oct 2015.

[63] Han-Pang Chiu, Leslie Pack Kaelbling, and Tomas Lozano-Perez. Virtual training for multi-view object class recognition. *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007.

[64] François Chollet. Keras. https://github.com/fchollet/keras, 2015.

[65] Gerard Pons-Moll Christoph Lassner and Peter V. Gehler. A generative model of people in clothing. *CoRR*, abs/1705.04098, 2017.

[66] Ondrej Chum and Andrew Zisserman. An exemplar model for learning object classes. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, jun 2007.

[67] Cmu. Carnegie-Mellon Mocap Database.

[68] David Cohen-Steiner and Frank Da. A greedy Delaunay-based surface reconstruction algorithm. *Vis. Comput.*, 20(1):4–16, apr 2004.

[69] Ronald R. Coifman and Stéphane Lafon. Diffusion maps. *Applied and Computational Harmonic Analysis*, 21(1):5–30, 2006.

[70] Jamie Condliffe. A key piece of self-driving cars is about to get a lot cheaper. *MIT Techlogy Review*, 2017.

[71] Microsoft Corporation. Kinect for windows sdk beta programming guide beta 1 draft version 1.1. Technical report, Microsoft, 2011.

[72] A Criminisi, I Reid, and A Zisserman. Single view metrology. *Int. J. Comput. Vision*, 40(2):123–148, Nov 2000.

[73] Antonio Criminisi, Andrew Zisserman, Luc van Gool, Simon Bramble, and David Compton. A new approach to obtain height measurements from video. In *Proc. of SPIE*, volume 3576, Jan 1998.

[74] A. B. Cua, K. P. Wilhelm, and H. I. Maibach. Elastic properties of human skin: relation to age, sex, and anatomical region. *Archives of Dermatological Research*, 282(5):283–288, aug 1990.

[75] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 886–893. IEEE, 2005.

[76] N. Daniell, T. Olds, and G. Tomkinson. Technical note: Criterion validity of whole body surface area equations: a comparison using 3D laser scanning. *Am J Phys Anthropol*, 148(1):148–55, 2012.

[77] Antitza Dantcheva, Petros Elia, and Arun Ross. What else does your biometric data reveal? A survey on soft biometrics. *IEEE Transactions on Information Forensics and Security*, 11(3):441–467, Mar 2016.

[78] Antitza Dantcheva, Carmelo Velardo, Angela D'angelo, and Jean-Luc Dugelay. Bag of soft biometrics for person identification: New trends and challenges. *Mutimedia Tools and Applications, Springer, October 2010*, 51(2):739–777, 2011.

[79] Parth Rajesh Desai, Pooja Nikhil Desai, Komal Deepak Ajmera, and Khushbu Mehta. A review paper on oculus rift-a virtual. *International Journal of Engineering Trends and Technology (IJETT)*, 13(4):175–179, 2014.

[80] Changxing Ding and Dacheng Tao. A comprehensive survey on pose-invariant face recognition. *ACM Trans. Intell. Syst. Technol.*, 7(3):37:1–37:42, February 2016.

[81] Ke Ding and Yun-Hui Liu. Sphere image for 3-d model retrieval. *IEEE Transactions on Multimedia*, 16(5):1369–1376, aug 2014.

[82] Frederick Drimmer. *Born Different: Amazing Stories Of Very Special People*. Bantam Books, New York, 1991.

[83] D Du Bois and E F Du Bois. A formula to estimate the approximate surface area if height and weight be known, 1916. *Nutrition*, 5(5):303–311, 1989.

[84] J V Durnin and J Womersley. Body fat assessed from total body density and its estimation from skinfold thickness: measurements on 481 men and women aged from 16 to 72 years. *The British journal of nutrition*, 32(1):77–97, jul 1974.

[85] Asi Elad Elbaz and Ron Kimmel. On bending invariant signatures for surfaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(10):1285–1295, 2003.

[86] Claire C. Gordon et al. Anthropometric survey of u.s. army personnel: summary statistics, interim report for 1998. Technical report, UNITED STATES ARMY NATICK, 1989.

[87] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, June 2010.

[88] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. LIBLINEAR: A library for large linear classification. *The Journal of Machine Learning Research*, 9:1871–1874, 2008.

[89] A Felici, J Verweij, and A Sparreboom. Dosing strategies for anticancer drugs: the good, the bad and body-surface area. *Eur. J. Cancer*, 38(13):1677–1684, sep 2002.

[90] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, Sept 2010.

[91] Pedro Felzenszwalb, David McAllester, and Deva Ramanan. A discriminatively trained, multiscale, deformable part model. *26th IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 1–8, 2008.

[92] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, volume 2, pages II–264–II–271. IEEE Comput. Soc, 2003.

[93] M P Hayes M Fink and N Soni. *Classic Papers in Critical Care*. Springer, 2008.

[94] U.S. Preventive Services Task Force. Screening for obesity in adults: recommendations and rationale. *Ann. Intern. Med.*, 139(11):930–932, dec 2003.

[95] US Preventive Services Task Force. Screening for obesity in children and adolescents: Us preventive services task force recommendation statement. *Pediatrics*, 125(2):361–367, 2010.

[96] Oren Freifeld and Michael J Black. Lie Bodies: A manifold representation of 3d human shape. In Fitzgibbon et al., editor, *ECCV (1)*, volume 7572 of *Lecture Notes in Computer Science*, pages 1–14. Springer, 2012.

[97] Oren Freifeld, Søren Hauberg, and Michael J Black. Model Transport: Towards scalable transfer learning on manifolds. In *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '14, pages 1378–1385. IEEE Computer Society, 2014.

[98] Thomas Funkhouser, Patrick Min, Michael Kazhdan, Joyce Chen, Alex Halderman, David Dobkin, and David Jacobs. A search engine for 3D models. *ACM Transactions on Graphics*, 22(1):83–105, jan 2003.

[99] D Gallagher, M Visser, D Sepulveda, R N Pierson, T Harris, and S B Heymsfield. How useful is body mass index for comparison of body fatness across age, sex, and ethnic groups? *Am. J. Epidemiol.*, 143(3):228–239, feb 1996.

[100] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576*, aug 2015.

[101] E. A. Gehan and S. L. George. Estimation of human body surface area from height and weight. *Cancer Chemother Rep*, 54(4):225–235, Aug 1970.

[102] Andrew. Gelman and Jennifer Hill. *Data Analysis using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, 2007.

[103] Ian Goodfellow. Nips 2016 tutorial: Generative adversarial networks. *arXiv preprint arXiv:1701.00160*, 2016.

[104] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014.

[105] C Gordon, D Webb, and S Wolpert. Isospectral plane domains and surfaces via Riemannian orbifolds. *Inventiones Mathematicae*, 110(1):1–22, 1992.

[106] T.M. Graber. Anthropometry of the head and face in medicine. *American Journal of Orthodontics*, 82(5):438, nov 1982.

[107] U. Grenander. *Pattern Analysis*, volume 24 of *Applied Mathematical Sciences*. Springer New York, New York, NY, 1978.

[108] Ulf Grenander and Michael I Miller. Representations of knowledge in complex systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 549–603, 1994.

[109] Ulf Grenander and Michael I Miller. Computational anatomy: An emerging discipline. *Quarterly of Applied Mathematics*, 56(4):617–694, 1998.

[110] Ulf Grenander and Michael I Miller. *Pattern Theory: From Representation to Inference*. Oxford University Press, UK, 2006.

[111] Hu Han, Charles Otto, Xiaoming Liu, and Anil K. Jain. Demographic estimation from face images: Human vs. machine performance. *IEEE Trans. Pattern Anal. Mach. Intell.*, 37(6):1148–1161, 2015.

[112] J. Han, L. Shao, D. Xu, and J. Shotton. Enhanced computer vision with Microsoft Kinect sensor: a review. *IEEE Transactions on Cybernetics*, 43(5):1318–1334, Oct 2013.

[113] T. S. Han, C. E. Morrison, and M. E. Lean. Age and health indications assessed by silhouette photographs. *Eur J Clin Nutr*, 53(8):606–611, Aug 1999.

[114] R.˜I. Hartley and A Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521540518, second edition, 2004.

[115] N Hasler and C Stoll. A statistical model of human pose and body shape. *Computer Graphics*, 28(2):1–10, 2009.

[116] Nils Hasler, Carsten Stoll, Bodo Rosenhahn, Thorsten Thormählen, and Hans-Peter Seidel. Estimating body shape of dressed humans. *Computers & Graphics*, 33(3):211–216, 2009.

[117] George B Haycock, George J Schwartz, and David H Wisotsky. Geometric method for measuring body surface area: A height-weight formula validated in infants, children, and adults. *The Journal of Pediatrics*, 93(1):62–66, 1978.

[118] M. Hazewinkel. *Encyclopaedia of Mathematics*. Number 5 in Encyclopaedia of Mathematics. Springer Netherlands, 2013.

[119] Xuming He, R. S. Zemel, and M. A. Carreira-Perpinan. Multiscale conditional random fields for image labeling. In *Proceedings of the 2004 IEEE Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, volume 2, pages II–695–II–702 Vol.2, June 2004.

[120] Carlos Hernandez, Francis Schmitt, and Roberto Cipolla. Silhouette coherence for camera calibration under circular motion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(2):343–349, 2007.

[121] C Daniel Herrera, Juho Kannala, and Janne Heikkilä. Accurate and practical calibration of a depth and color camera pair. In *Proceedings of the 14th International Conference on Computer Analysis of Images and Patterns - Volume Part II*, ICAIP'11, pages 437–445, Berlin, Heidelberg, 2011. Springer-Verlag.

[122] S Hettiaratchy and R Papini. Initial management of a major burn: Vol II: assessment and resuscitation. *BMJ*, 329(7457):101–103, Jul 2004.

[123] Aaron Heysse. *A Survey of Modern Rendering Techniques*. PhD thesis, ETH Zurich, 2012.

[124] Hugo Hidalgo, Sonia Sosa León, and Enrique Gómez-Treviño. Application of the kernel method to the inverse geosounding problem. *Neural Networks : The Official Journal of the International Neural Network Society*, 16(3-4):349–53, jan 2003.

[125] Masaki Hilaga, Yoshihisa Shinagawa, Taku Kohmura, and Tosiyasu L. Kunii. Topology matching for fully automatic similarity estimation of 3D shapes. In *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques - SIGGRAPH '01*, pages 203–212, New York, New York, USA, aug 2001. ACM Press.

[126] Geoffrey E Hinton, Simon Osindero, and Yee-Whye Teh. A fast learning algorithm for deep belief nets. *Neural Computation*, 18(7):1527–54, 2006.

[127] David A. Hirshberg, Matthew Loper, Eric Rachlin, and Michael J. Black. Coregistration: Simultaneous alignment and modeling of articulated 3D shape. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 7577 LNCS(PART 6):242–255, 2012.

[128] Dirk Holz and Sven Behnke. Fast range image segmentation and smoothing using approximate surface reconstruction and region growing. In Sukhan Lee, Hyungsuck Cho, Kwang-Joon Yoon, and Jangmyung Lee, editors, *Intelligent Autonomous Systems 12*, volume 194 of *Advances in Intelligent Systems and Computing*, pages 61–73. Springer Berlin Heidelberg, 2013.

[129] S Holzer, R B Rusu, M Dixon, S Gedikli, and N Navab. Adaptive neighborhood selection for real-time surface normal estimation from organized point cloud data using

integral images. In *Intelligent Robots and Systems (IROS), 2012 IEEE/RSJ International Conference on*, pages 2684–2689, 2012.

[130] S D Hsieh, H Yoshinaga, and T Muto. Waist-to-height ratio, a simple and practical index for assessing central fat distribution and metabolic risk in Japanese men and women. *International Journal of Obesity and Related Metabolic Disorders : Journal of the International Association for the Study of Obesity*, 27(5):610–6, may 2003.

[131] Masatsugu Ichino and Yasushi Yamazaki. Soft biometrics and its application to security and business. In *Proceedings - 2013 International Conference on Biometrics and Kansei Engineering, ICBAKE 2013*, pages 314–319, 2013.

[132] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of The 32nd International Conference on Machine Learning*, pages 448–456, 2015.

[133] J. Sivic and A. Zisserman. Video Google: a text retrieval approach to object matching in videos. In *Proceedings Ninth IEEE International Conference on Computer Vision*, pages 1470–1477 vol.2. IEEE, 2003.

[134] A S Jackson, P R Stanforth, J Gagnon, T Rankinen, A S Leon, D C Rao, J S Skinner, C Bouchard, and J H Wilmore. The effect of sex, age and race on estimating percentage body fat from body mass index: The Heritage Family Study. *International Journal of Obesity and Related Metabolic Disorders : Journal of the International Association for the Study of Obesity*, 26(6):789–96, jun 2002.

[135] Anil K Jain, Sarat C Dass, Karthik Nandakumar, Karthik N, and N Karthik. Soft biometric traits for personal recognition systems. In *Proceedings of International Conference on Biometric Authentication, Hong Kong*, pages 731–738, 2004.

[136] P Jain, B Kulis, and K Grauman. Fast image search for learned metrics. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2008.

[137] N Jander, C Gohlke-Barwolf, E Bahlmann, E Gerdts, K Boman, J B Chambers, K Egstrup, C A Nienaber, T R Pedersen, S Ray, A B Rossebø, R Willenheimer, R P Kienzle, K Wachtell, F J Neumann, and J Minners. Indexing aortic valve area by body surface area increases the prevalence of severe aortic stenosis. *Heart*, 100(1):28–33, jan 2014.

[138] A.E. Johnson and M. Hebert. Using spin images for efficient object recognition in cluttered 3D scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(5):433–449, May 1999.

[139] Jonathan Masci, Davide Boscaini, Michael M. Bronstein, and Pierre Vandergheynst. Geodesic convolutional neural networks on Riemannian manifolds. In *Computer Vision Workshop (ICCVW), 2015 IEEE International Conference on*, pages 37–45, 2015.

[140] Mark Kac. Can one hear the shape of a drum? *The American Mathematical Monthly*, 73(4):1, Apr 1966.

[141] S H Kaplan and Stanley H Kaplan Educational Center. *Pharmacology*. Basic Medical Science Notes. S.H. Kaplan Educational Center Limited, 1988.

[142] P T Katzmarzyk and W R Leonard. Climatic influences on human body size and proportions: ecological adaptations and secular trends. *American Journal of Physical Anthropology*, 106(4):483–503, aug 1998.

[143] Michael Kazhdan, Matthew Bolitho, and Hugues Hoppe. Poisson surface reconstruction. In *Proceedings of the Fourth Eurographics Symposium on Geometry Processing*,

SGP '06, pages 61–70, Aire-la-Ville, Switzerland, Switzerland, 2006. Eurographics Association.

[144] Michael Kazhdan, Thomas Funkhouser, and Szymon Rusinkiewicz. Rotation invariant spherical harmonic representation of 3D shape descriptors. In *Proceedings of the 2003 Eurographics/ACM SIGGRAPH Symposium on Geometry Processing*, SGP '03, pages 156–164, Aire-la-Ville, Switzerland, Switzerland, 2003. Eurographics Association.

[145] Michael M. Kazhdan. *Shape Representations and Algorithms for Three-Dimensional Model Retrieval*. PhD thesis, Princeton, Jan 2004.

[146] Ancel Keys, Flaminio Fidanza, Martti J. Karvonen, Noboru Kimura, and Henry L. Taylor. Indices of relative weight and obesity. *Journal of Chronic Diseases*, 25(6-7):329–343, jul 1972.

[147] D. P Kingma and M. Welling. Auto-encoding variational Bayes. *ArXiv e-prints*, December 2013.

[148] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[149] I. Kokkinos, M. M. Bronstein, R. Litman, and A. M. Bronstein. Intrinsic shape context descriptors for deformable shapes. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 159–166, June 2012.

[150] Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques - Adaptive Computation and Machine Learning*. The MIT Press, 2009.

[151] Nir Y Krakauer and Jesse C Krakauer. A new body shape index predicts mortality hazard independently of body mass index. *PloS ONE*, 7(7):e39504, 2012.

[152] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, U. Toronto, 2009.

[153] A. Kushal, C. Schmid, and J. Ponce. Flexible object models for category-level 3D object recognition. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, June 2007.

[154] Young H Kwon and Niels da Vitoria Lobo. Age classification from facial images. *Computer Vision and Image Understanding*, 74(1):1–21, apr 1999.

[155] John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

[156] Andreas Lanitis. Evaluating the performance of face-aging algorithms. In *8th IEEE International Conference on Automatic Face Gesture Recognition*, pages 1–6, Sept 2008.

[157] O Lapid. Measuring the surface area of surgical lesions. *Ann Plast Surg*, 58(6):706; author reply 706, jun 2007.

[158] Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle, and Ole Winther. Autoencoding beyond pixels using a learned similarity metric. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, pages 1558–1566. JMLR.org, 2016.

[159] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1(4):541–551, Dec 1989.

[160] Crystal Man Ying Lee, Rachel R Huxley, Rachel P Wildman, and Mark Woodward. Indices of abdominal obesity are better discriminators of cardiovascular risk factors than BMI: a meta-analysis. *Journal of clinical epidemiology*, 61(7):646–53, Jul 2008.

[161] Sung-Hee Lee, Eftychios Sifakis, and Demetri Terzopoulos. Comprehensive biomechanical modeling and simulation of the upper body. *ACM Transactions on Graphics*, 28(4):1–17, Aug 2009.

[162] Young-A Lee, Susan P Ashdown, and Ann C Slocum. Measurement of surface area of 3-D body scans to assess the effectiveness of hats for sun protection. *Family and Consumer Sciences Research Journal*, 34(4):366–385, 2006.

[163] Joel Z Leibo, Jim Mutch, and Tomaso Poggio. Why the brain separates face recognition from object recognition. *Advances in Neural Information Processing Systems (NIPS)*, pages 1–9, 2011.

[164] Bruno Levy. Laplace-Beltrami Eigenfunctions: Towards an algorithm that understand geometry. *IEEE International Conference on Shape Modeling and Applications, inv ited talk*, 2006.

[165] B Li, T Schreck, A Godil, M Alexa, T Boubekeur, B Bustos, J Chen, M Eitz, and T Furuya. SHREC12 Track: Sketch-based 3D shape retrieval. In *Eurographics Workshop on 3D Object Retrieval 2012 (3DOR 2012)*, pages 109–118, 2012.

[166] Bo Li, Afzal Godil, and Henry Johan. Hybrid shape descriptor and meta similarity generation for non-rigid and partial 3D model retrieval. *Multimedia Tools and Applications*, 72(2):1531–1560, apr 2013.

[167] Bo et al. Li. A comparison of 3D shape retrieval methods based on a large-scale

benchmark supporting multimodal queries. *Computer Vision and Image Understanding*, 131:1–27, feb 2015.

[168] Chunyuan Li and A. Ben Hamza. Spatially aggregating spectral descriptors for nonrigid 3D shape retrieval: A comparative survey. *Multimedia Systems*, 20(3):253–281, 2014.

[169] Stan Z. Li and Anil Jain. *Encyclopedia of Biometrics*. Springer US, Boston, MA, 2009.

[170] Stan Z. Li and Anil K. Jain. *Handbook of Face Recognition*. Springer Publishing Company, Incorporated, 2nd edition, 2011.

[171] Z. Lian, A. Godil, T. Fabry, T. Furuya, J. Hermans, R. Ohbuchi, C. Shu, D. Smeets, P. Suetens, D. Vandermeulen, and S. Wuhrer. SHREC'10 Track: Non-rigid 3D shape retrieval. In Mohamed Daoudi and Tobias Schreck, editors, *Eurographics Workshop on 3D Object Retrieval*. The Eurographics Association, 2010.

[172] J Liebelt, C Schmid, and K Schertler. Viewpoint-independent object class detection using 3D feature maps. *IEEE Conference on Computer Vision and Pattern Recognition (2008)*, 64(6):1–8, 2008.

[173] Olof Lindberg, Mark Walterfang, Jeffrey C L Looi, Nikolai Malykhin, Per Ostberg, Bram Zandbelt, Martin Styner, Beatriz Paniagua, Dennis Velakoulis, Eva Orndahl, and Lars-Olof Wahlund. Hippocampal shape analysis in Alzheimer's disease and frontotemporal lobar degeneration subtypes. *Journal of Alzheimer's Disease : JAD*, 30(2):355–65, jan 2012.

[174] Joakim Lindblad and Ingela Nyström. Surface area estimation of digitized 3D objects using local computations. In Achille Braquelaire, Jacques-Olivier Lachaud, and Anne Vialard, editors, *Discrete Geometry for Computer Imagery*, volume 2301 of *Lecture Notes in Computer Science*, pages 267–278. Springer Berlin Heidelberg, 2002.

[175] Yaron Lipman and Thomas Funkhouser. Möbius voting for surface correspondence. *ACM Transactions on Graphics*, 28(3):1, Jul 2009.

[176] R. Litman and A. M. Bronstein. Learning spectral descriptors for deformable shape correspondence. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(1):171–180, 2014.

[177] Roee Litman, Alex Bronstein, Michael Bronstein, and Umberto Castellani. Supervised learning of bag-of-features shape descriptors using sparse coding. In *Computer Graphics Forum*, volume 33, pages 127–136. Wiley Online Library, 2014.

[178] Qiong Liu. A survey of recent view-based 3D model retrieval methods. *CoRR*, abs/1208.3670:15, aug 2012.

[179] W. Liu, Y. Zhang, S. Tang, J. Tang, R. Hong, and J. Li. Accurate estimation of human body orientation from RGB-D sensors. *IEEE Transactions on Cybernetics*, 43(5):1442–1452, Oct 2013.

[180] Yang Liu, Balakrishnan Prabhakaran, and Xiaohu Guo. Point-based manifold harmonics. *IEEE Transactions on Visualization and Computer Graphics*, 18(10):1693–1703, 2012.

[181] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, Dec 2015.

[182] E H Livingston and S Lee. Body surface area prediction in normal-weight and obese patients. *Am. J. Physiol. Endocrinol. Metab.*, 281(3):E586—-591, sep 2001.

[183] Matthew M. Loper and Michael J. Black. OpenDR: An approximate and differentiable renderer. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors,

Computer Vision – ECCV 2014, pages 154–169, Cham, 2014. Springer International Publishing.

[184] Matthew M. Loper, Naureen Mahmood, and Michael J. Black. MoSh: Motion and shape capture from sparse markers. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 33(6):220:1–220:13, November 2014.

[185] William E Lorensen and Harvey E Cline. Marching Cubes: a high resolution 3d surface construction algorithm. *SIGGRAPH Comput. Graph.*, 21(4):163–169, aug 1987.

[186] David G Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2):91–110, Nov 2004.

[187] Jun-Ming Lu and Mao-Jiun J Wang. Automated anthropometric data collection using 3D whole body scanners. *Expert Systems with Applications*, 35(1–2):407–414, 2008.

[188] Teghan Lucas and Maciej Henneberg. Comparing the face to the body, which is better for identification? *International Journal of Legal Medicine*, 130(2):533–540, 2016.

[189] Meysam Madadi, Sergio Escalera, Jordi Gonzalez, F. Xavier Roca, and Felipe Lumbreras. Multi-part body segmentation based on depth maps for soft biometry analysis. *Pattern Recognition Letters*, 56:14–21, 2015.

[190] Patrick Mair, Felix Schoenbrodt, and Rand Wilcox. *WRS2: Wilcox Robust Estimation and Testing*, 2016.

[191] Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, and Ian Goodfellow. Adversarial autoencoders. In *International Conference on Learning Representations*, 2016.

[192] S Manay, B Hong, a Yezzi, and S Soatto. Integral invariant signatures. *Computer Vision - ECCV 2004*, 3024:87–99, 2004.

[193] David Marr. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. The MIT Press, 2010.

[194] J Matas, O Chum, M Urban, and T Pajdla. Robust wide-baseline stereo from maximally stable extremal regions. *In British Machine Vision Conference*, pages 384–393, 2002.

[195] Walter Maurel, Daniel Thalmann, Yin Wu, and Nadia Magnenat Thalmann. *Biomechanical Models for Soft Tissue Simulation*. Springer Berlin Heidelberg, Berlin, Heidelberg, 1998.

[196] Jr. H P McKean and I M Singer. Curvature and the eigenvalues of the Laplacian. *Journal of Differential Geometry*, 100(1-2):43–69, 1967.

[197] R. Meghana. User experience guidelines for Intel realsense applications: There's an app for that. Technical report, Intel Developer Zone: https://software.intel.com/en-us/articles/implementing-user-experience-guidelines-in-intel-realsense-applications, 2016.

[198] Z Mei, L M Grummer-Strawn, A Pietrobelli, A Goulding, M I Goran, and W H Dietz. Validity of body mass index compared with other body-composition screening indexes for the assessment of body fatness in children and adolescents. *Am. J. Clin. Nutr.*, 75(6):978–985, jun 2002.

[199] Microsoft Research. Kinect for Windows SDK beta. Technical report, MSR, 2011.

[200] R. D. Mosteller. Simplified calculation of body-surface area. *N. Engl. J. Med.*, 317(17):1098, Oct 1987.

[201] S. Motiian, Q. Jones, S. M. Iranmanesh, and G. Doretto. Few-shot adversarial domain adaptation. *ArXiv e-prints*, November 2017.

[202] Saeid Motiian, Marco Piccirilli, Donald A. Adjeroh, and Gianfranco Doretto. Unified deep supervised domain adaptation and generalization. In *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[203] B. Triggs N. Dalal. Histograms of oriented gradients for human detection. In *Proceeding of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005.

[204] V Nair and G E Hinton. Rectified linear units improve restricted Boltzmann machines. In *ICML'2010*, 2010.

[205] Pedro J. Navarro, Carlos Fernndez, Ral Borraz, and Diego Alonso. A machine learning approach to pedestrian detection for autonomous vehicles using high-definition 3d range data. *Sensors*, 17(1), 2017.

[206] C V Nguyen, S Izadi, and D Lovell. Modeling Kinect Sensor Noise for Improved 3D Reconstruction and Tracking. In *3D Imaging, Modeling, Processing, Visualization and Transmission (3DIMPVT), 2012 Second International Conference on*, pages 524–530, oct 2012.

[207] Ngoc Hung Nguyen and Richard Hartley. Height measurement for humans in motion using a camera: A comparison of different methods. In *2012 International Conference on Digital Image Computing Techniques and Applications (DICTA)*, pages 1–8. IEEE, Dec 2012.

[208] Mark S. Nixon, Paulo L. Correia, Kamal Nasrollahi, Thomas B. Moeslund, Abdenour Hadid, and Massimo Tistarelli. On soft biometrics. *Pattern Recognition Letters*, 68:218 – 230, 2015. Special Issue on Soft Biometrics.

[209] G. L. Oliveira, A. Valada, C. Bollen, W. Burgard, and T. Brox. Deep learning for hu-

man part discovery in images. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1634–1641, May 2016.

[210] Robert Osada, Thomas Funkhouser, Bernard Chazelle, and David Dobkin. Shape distributions. *ACM Transactions on Graphics*, 21(4):807–832, 2002.

[211] M Ovsjanikov, A M Bronstein, M M Bronstein, and L J Guibas. Shape Google: A computer vision approach to isometry invariant shape retrieval. In *Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on*, pages 320–327, sep 2009.

[212] Eric Paquet, Marc Rioux, Anil Murching, and Thumpudi Naveen. Description of shape information for 2-D and 3-D objects. *Signal Processing: Image Communication*, 16(1-2):103–122, sep 2000.

[213] Devi Parikh and Kristen Grauman. Relative attributes. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 503–510, 2011.

[214] S Parsons. *Pharmaceutical Calculations*. Parsons Printing Press, 2012.

[215] Niklas Peinecke, Franz Erich Wolter, and Martin Reuter. Laplace spectra as fingerprints for image recognition. *CAD Computer Aided Design*, 39(6):460–476, 2007.

[216] X. Peng, B. Usman, N. Kaushik, J. Hoffman, D. Wang, and K. Saenko. VisDA: The Visual Domain Adaptation Challenge. *ArXiv e-prints*, October 2017.

[217] F Perronnin and C Dance. Fisher kenrels on visual vocabularies for image categorizaton. In *2006 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2006.

[218] Florent Perronnin, Jorge Sánchez, and Thomas Mensink. Improving the Fisher kernel for large-scale image classification. In *In Proceeding of European Conference of Computer Vision,*, page 3, 2010.

[219] Mary A. Peterson and Gillian Rhodes. *Perception of Faces, Objects, and Scenes: Analytic and Holistic Processes*. Oxford University Press, 2003, 2006.

[220] Bui Tuong Phong. Illumination for computer generated pictures. *Communications of the ACM*, 18(6):311–317, 1975.

[221] M. Piccirilli, G. Doretto, A. Ross, and D. Adjeroh. A mobile structured light system for 3d face acquisition. *IEEE Sensors Journal*, 16(7):1854–1855, April 2016.

[222] Marco Piccirilli, Gianfranco Doretto, and Donald Adjeroh. A framework for analyzing the whole body surface area from a single view. *PLOS ONE*, 12(1):e0166749, Jan 2017.

[223] D Pickup, X Sun, and et. al. Shrec'14 track: Shape retrieval of non-rigid 3d human models. In *Proceedings of the 7th Eurographics Workshop on 3D Object Retrieval*, 3DOR '15, pages 101–110, Aire-la-Ville, Switzerland, Switzerland, 2014. Eurographics Association.

[224] Sébastien Piérard, Damien Leroy, Jean-Frédéric Hansen, and Marc Van Droogenbroeck. Estimation of human orientation in images captured with a range camera. In *Proceedings of the 13th International Conference on Advanced Concepts for Intelligent Vision Systems*, ACIVS'11, pages 519–530, Berlin, Heidelberg, 2011. Springer-Verlag.

[225] Jose Pinheiro, Douglas Bates, Saikat DebRoy, Deepayan Sarkar, and R Core Team. *nlme: Linear and Nonlinear Mixed Effects Models*, 2016. R package version 3.1-137.

[226] Ulrich Pinkall and Konrad Polthier. Computing discrete minimal surfaces and their conjugates. *EXPERIMENTAL MATHEMATICS*, 2:15–36, 1993.

[227] Gerard Pons-Moll, Javier Romero, Naureen Mahmood, and Michael J. Black. Dyna: A model of dynamic human shape in motion. *ACM Transactions on Graphics, (Proc. SIGGRAPH)*, 34(4):120:1–120:14, August 2015.

[228] Thomas Porter and Tom Duff. Compositing digital images. *ACM SIGGRAPH Computer Graphics*, 18(3):253–259, jul 1984.

[229] M H Protter. Can One Hear the Shape of a Drum? Revisited. *SIAM Review*, 29(2):185–197, jun 1987.

[230] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3D classification and segmentation. In *Proceeding of the IEEE Conference of Computer Vision and Pattern Recognition*, 2016.

[231] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008. ISBN 3-900051-07-0.

[232] R Development Core Team and R Core Team. R: A Language and Environment for Statistical Computing, 2014.

[233] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *http://arxiv. org/abs/1511*, 6434, 2015.

[234] Syed Ashiqur Rahman and Donald Adjeroh. Surface-based body shape index and Its relationship with all-Causeause mortality. *PloS one*, 10(12):e0144639, 2015.

[235] D. Ramanan. Using segmentation to verify object hypotheses. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, June 2007.

[236] Lillian J Ratliff, Samuel A Burden, and S Shankar Sastry. Characterization and computation of local nash equilibria in continuous games. In *Communication, Control, and Computing (Allerton), 2013 51st Annual Allerton Conference on*, pages 917–924. IEEE, 2013.

[237] Dan Raviv, Michael M. Bronstein, Alexander M. Bronstein, Ron Kimmel, and Nir Sochen. Affine-invariant diffusion geometry for the analysis of deformable 3D shapes. In *2011 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2361–2367, 2011.

[238] D. A. Reid, M. S. Nixon, and S. V. Stevenage. Soft biometrics; human identification using comparative descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(6):1216–1228, June 2014.

[239] D. A. Reid, S. Samangooei, C. Chen, M. S. Nixon, and A. Ross. Soft biometrics for surveillance: An overview. *Handbook of Statistics*, 31:327–352, 2013.

[240] Martin Reuter. *Laplace Spectra for Shape Recognition*. PhD thesis, University of Hanover, 2006.

[241] Martin Reuter, Franz-Erich Wolter, and Niklas Peinecke. Laplace-Beltrami spectra as 'Shape-DNA' of surfaces and solids. *Computer-Aided Design*, 38(4):342–366, 2006.

[242] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic back-propagation and approximate inference in deep generative models. *arXiv preprint arXiv:1401.4082*, 2014.

[243] Karl Ricanek and Benjamin Barbour. What are soft biometrics and how can they be used? *Computer*, 44(9):106–108, 2011.

[244] K M Robinette, H Daanen, and E Paquet. The CAESAR project: a 3-D surface anthropometry survey. In *3-D Digital Imaging and Modeling, 1999. Proceedings. Second International Conference on*, pages 380–386, 1999.

[245] R. Tyrrell Rockafellar and Roger J. B. Wets. *Variational Analysis*, volume 317 of *Grundlehren der mathematischen Wissenschaften*. Springer Berlin Heidelberg, Berlin, Heidelberg, 1998.

[246] Dror M. Rom. A sequentially rejective test procedure based on a modified Bonferroni inequality. *Biometrika*, 77(3):663, sep 1990.

[247] S Rosenberg. *The Laplacian on a Riemannian Manifold*. Cambridge University Press, Cambridge, UK, 1997.

[248] Scott D Roth. Ray casting for modeling solids. *Computer Graphics and Image Processing*, 18(2):109–144, feb 1982.

[249] Sam T. Roweis and Lawrence K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

[250] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252, sep 2015.

[251] Olga Russakovsky and Li Fei-Fei. Attribute learning in large-scale datasets. In *Lecture Notes in Computer Science*, volume 6553 LNCS, pages 1–14, 2012.

[252] S J Russell and P Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall Series in Artificial Intelligence. Prentice Hall, 2010.

[253] J J Sacco, J Botten, F Macbeth, A Bagust, and P Clark. The average body surface area of adult cancer patients in the UK: A multicentre retrospective study. *PLoS ONE*, 5(1):e8933, 2010.

[254] G Salvendy. *Handbook of Human Factors and Ergonomics*. Wiley, 2012.

[255] L B Sardinha, A M Silva, C S Minderico, and P J Teixeira. Effect of body surface area calculations on body fat estimates in non-obese and obese subjects. *Physiol Meas*, 27(11):1197–1209, nov 2006.

[256] A Sarría, L A García-Llop, L A Moreno, J Fleta, M P Morellón, and M Bueno. Skinfold thickness measurements are better predictors of body fat percentage than body mass index in male Spanish children and adolescents. *European journal of clinical nutrition*, 52(8):573–6, aug 1998.

[257] S Savarese and L Fei-fei. 3d generic object categorization, localization and pose estimation. In *2007 IEEE 11th International Conference on Computer Vision*, 2007.

[258] Silvio Savarese and Li Fei-Fei. View synthesis for recognizing unseen poses of object classes. *Lecture Notes in Computer Science*, 5304 LNCS(PART 3):602–615, 2008.

[259] M Sawyer and M J Ratain. Body surface area as a determinant of pharmacokinetics and drug dosing. *Invest New Drugs*, 19(2):171–177, may 2001.

[260] W. J. Scheirer, N. Kumar, V. N. Iyer, P. N. Belhumeur, and T. E. Boult. How reliable are your visual attributes? In Ioannis Kakadiaris, Walter J. Scheirer, and Laurence G.

Hassebrook, editors, *Proceedings of the SPIE, Volume 8712, id. 87120Q 12 pp. (2013).*, volume 8712, page 87120Q, May 2013.

[261] Walter J. Scheirer, Neeraj Kumar, Karl Ricanek, Peter N. Belhumeur, and Terrance E. Boult. Fusing with context: A Bayesian approach to combining descriptive attributes. In *2011 International Joint Conference on Biometrics, IJCB 2011*, 2011.

[262] Harald J Schneider, Nele Friedrich, Jens Klotsche, Lars Pieper, Matthias Nauck, Ulrich John, Marcus Dörr, Stephan Felix, Hendrik Lehnert, David Pittrow, Sigmund Silber, Henry Völzke, Günter K Stalla, Henri Wallaschofski, and Hans-Ulrich Wittchen. The predictive value of different measures of obesity for incident cardiovascular events and mortality. *The Journal of Clinical Endocrinology and Metabolism*, 95(4):1777–85, apr 2010.

[263] Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Comput.*, 10(5):1299–1319, July 1998.

[264] Seok-Han Lee, Tae-Eun Kim, and Jong-Soo Choi. A single-view based framework for robust estimation of heights and positions of moving people. In *2010 Digest of Technical Papers International Conference on Consumer Electronics (ICCE)*, pages 503–504. IEEE, Jan 2010.

[265] Ken Shoemake. Animating rotation with quaternion curves. *SIGGRAPH Comput. Graph.*, 19(3):245–254, July 1985.

[266] Jamie Shotton, Andrew Fitzgibbon, Mat Cook, Toby Sharp, Mark Finocchio, Richard Moore, Alex Kipman, and Andrew Blake. Real-time human pose recognition in parts from single depth images. In *In Proc. of Computer Vision and Pattern Recognition, 2011*, 2011.

[267] Martin Simonovsky and Nikos Komodakis. Dynamic edge-conditioned filters in convolutional neural networks on graphs. *arXiv preprint arXiv:1704.02901*, 2017.

[268] T. A. B. Snijders and R. J. (Roel J.) Bosker. *Multilevel Analysis : an Introduction to Basic and Advanced Multilevel Modeling*. Sage, 2012.

[269] Zheng Song, Meng Wang, Xian Sheng Hua, and Shuicheng Yan. Predicting occupation via human clothing and contexts. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1084–1091, 2011.

[270] Alex Sparreboom and Jaap Verweij. Paclitaxel pharmacokinetics, threshold models, and dosing strategies. *Journal of Clinical Oncology*, 21(14):2803–2804, 2003.

[271] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller. Striving for simplicity: The all convolutional net. *ArXiv e-prints*, December 2014.

[272] Valentina Staneva and Laurent Younes. Modeling and estimation of shape deformation for topology-preserving object tracking. *SIAM Journal on Imaging Sciences*, 7(1):427–455, 2014.

[273] David Zeltzer Steven M. Drucker. Intelligent camera control in a virtual environment. In *In Proceedings of Graphics Interface 94*, pages 190–199, 2016.

[274] Hang Su, Subhransu Maji, Evangelos Kalogerakis, and Erik Learned-Miller. Multi-view convolutional neural networks for 3D shape recognition. In *arXiv:1505.00880 [cs]*, 2015.

[275] Hao Su, Charles R. Qi, Yangyan Li, and Leonidas Guibas. Render for CNN: Viewpoint estimation in images using CNNs trained with rendered 3D model views. In *Proceedings of the IEEE International Conference on Computer Vision*, 2015.

[276] Jian Sun, Maks Ovsjanikov, and Leonidas Guibas. A concise and provably informative multi-scale signature based on heat diffusion. *Eurographics Symposium on Geometry Processing*, 28(5):1383–1392, 2009.

[277] Min Sun, Hao Su, Silvio Savarese, and Li Fei-Fei. A multi-view probabilistic model for 3D object classes. *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, 1:1247–1254, 2009.

[278] H. Sundar, D. Silver, N. Gagvani, and S. Dickinson. Skeleton based shape matching and retrieval. In *2003 Shape Modeling International.*, page 130. IEEE Computer Society, may 2003.

[279] J W H Tangelder and R C Veltkamp. A survey of content based 3D shape retrieval methods. *Shape Modeling Applications, 2004. Proceedings*, pages 145–156, 2004.

[280] Dacheng Tao, Xiaoou Tang, Xuelong Li, and Xindong Wu. Asymmetric bagging and random subspace for support vector machines-based relevance feedback in image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(7):1088–99, jul 2006.

[281] Michael J. Tarr and Heinrich H. Bülthoff. Is human object recognition better described by geon structural descriptions or by multiple views? Comment on Biederman and Gerhardstein (1993).

[282] Alexander Thomas, Vittorio Ferrari, Bastian Leibe, Tinne Tuytelaars, Bernt Schiele, and Luc Van Gool. Towards multi-view object class detection. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2:1589–1596, 2006.

[283] D.A.W. Thompson. *On Growth and Form*. Number v. 1 in Dover Books on Biology Series. Dover, 1942.

[284] Tijmen Tieleman and Geoffrey Hinton. Lecture 6.5 - rmsprop, 2012.

[285] Michal Tölgyessy and Peter Hubinský. The Kinect sensor in robotics education. In *In Proceedings of 2nd International Conference on Robotics in Education*, pages 143–146, 2010.

[286] Dorien Van Deun, Vincent Verhaert, Koen Buys, Bart Haex, and Jos Vander Sloten. Automatic generation of personalized human models based on body measurements. In *International Symposium on Digital Human Modeling edition:1*, volume 1, Lyon, June 2011. International Symposium on Digital Human Modeling.

[287] T Van Gestel, J A K Suykens, G Lanckriet, A Lambrechts, B De Moor, and J Vandewalle. Bayesian framework for least-squares support vector machine classifiers, Gaussian processes, and kernel Fisher discriminant analysis. *Neural Computation*, 14(5):1115–47, May 2002.

[288] A Vedaldi and K Lenc. MatConvNet – Convolutional Neural Networks for MATLAB.

[289] Andrea Vedaldi and Andrew Zisserman. Efficient additive kernels via explicit feature maps. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(3):480–492, 2012.

[290] Carmelo Velardo and Jean Luc Dugelay. Weight estimation from visual body appearance. In *2010 Fourth IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, 2010.

[291] Carmelo Velardo and Jean-luc Dugelay. What can computer vision tell you about your weight? In *2012 Proceedings of the 20th European Signal Processing Conference (EUSIPCO)*, pages 1980–1984, Aug 2012.

[292] J Verbraecken, P Van de Heyning, W De Backer, and L Van Gaal. Body surface area in normal-weight, overweight, and obese adults. A comparison study. *Metab. Clin. Exp.*, 55(4):515–524, apr 2006.

[293] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pages 1096–1103, New York, NY, USA, 2008. ACM.

[294] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 1, pages I–511–I–518. IEEE Comput. Soc, 2001.

[295] A. Weiss, D. Hirshberg, and M.J. Black. Home 3D body scans from noisy image and range data. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 1951–1958, 2011.

[296] B. Wikipedians. *3D Rendering*. PediaPress, 2016.

[297] R R Wilcox, H J Keselman, J Muska, and R Cribbie. Repeated measures ANOVA: Some new results on comparing trimmed means and means. *The British Journal of Mathematical and Statistical Psychology*, 53 (Pt 1):69–82, May 2000.

[298] John Winn and Jamie Shotton. *Markov Random Fields for Object Detection*, chapter 25, pages 389–404. MIT Press, 2011.

[299] Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. SUN database: Large-scale scene recognition from abbey to zoo. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3485–3492, 2010.

[300] Bing Xu, Naiyan Wang, Tianqi Chen, and Mu Li. Empirical evaluation of rectified activations in convolutional network. *arXiv preprint arXiv:1505.00853*, may 2015.

[301] X Yang Y Le. Tiny ImageNet Visual Recognition Challenge.

[302] Pingkun Yan, Saad M. Khan, and Mubarak Shah. 3D model based object class detection in an arbitrary view. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–6, 2007.

[303] X. Yang and Y. Tian. Super normal vector for human activity recognition with depth cameras. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(5):1028–1039, May 2017.

[304] C. Y. Yu, C. H. Lin, and Y. H. Yang. Human body surface area database and estimation formula. *Burns*, 36(5):616–629, Aug 2010.

[305] C Y Yu, Y H Lo, and W K Chiou. The 3D scanner for measuring body surface area: a simplified calculation in the Chinese adult. *Appl Ergon*, 34(3):273–278, may 2003.

[306] Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I*, pages 818–833, Cham, 2014. Springer International Publishing.

[307] Cha Zhang and Zhengyou Zhang. Calibration between depth and color sensors for commodity depth cameras. In *Multimedia and Expo (ICME), 2011 IEEE International Conference on*, pages 1–6, 2011.

[308] Xin Zhang, Yee-Hong Yang, Zhiguang Han, Hui Wang, and Chao Gao. Object class detection. *ACM Computing Surveys*, 46(1):1–53, oct 2013.

[309] Zhengyou Zhang. Flexible camera calibration by viewing a plane from unknown orientations. In *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, volume 1, pages 666–673 vol.1, 1999.

[310] Zhengyou Zhang. A flexible new technique for camera calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(11):1330–1334, 2000.

[311] Yunhong Zhou, Dennis Wilkinson, Robert Schreiber, and Rong Pan. Large-scale parallel collaborative filtering for the netflix prize. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 5034 LNCS:337–348, 2008.

[312] Zhi Zhou, Yue Wang, and Eam Khwang Teoh. *People Re-identification Based on Bags of Semantic Features*, pages 574–586. Springer International Publishing, Cham, 2015.

[313] Silvia Zuffi and Michael J. Black. The stitched puppet: A graphical model of 3D human shape and pose. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR 2015)*, pages 3537–3546, June 2015.

Appendix A

Multi-views Body Fat Percentage

A.1 Introduction

A longstanding question in computer vision concerns the representation of 3D shapes for recognition: should 3D shapes be represented with descriptors operating on their native 3D formats, such as voxel, grid or polygon mesh, or can they be effectively represented with view-based descriptors? [274] **This is one of most crucial debate in computer vision**.
In this work we face the task of human body BFP classification using traditional natural image descriptors. In Chapter 5 we leveraged **intrinsic** or pose/view-invariant descriptors. The present work, instead, considers a human body representation dependent on the coordinate system, and camera position.

For this task we use the recent findings in "Deep learning", which has the convolutional neural network (ConvNet) [159] architecture as the most known method. Convolutional Neural Networks (ConvNet) are very similar to ordinary Neural Networks (NNs): they are made up of neurons that have learnable weights and biases. Each neuron receives some inputs, performs a dot product and optionally follows it with a non-linear operation. The network still expresses

a single differentiable score function as in traditional Neural Networks. They still use a loss function (e.g., Softmax, Hinge loss, etc.) on the last (fully-connected) layer. ConvNet architectures, differently from NNs, make the explicit assumption that the inputs are images, which allows us to encode certain properties into the architecture. These make the forward function more efficient to implement and vastly reduce the number of parameters in the network.
ConvNets present many interesting properties, for a better explanation we refer to the survey [306]. ConvNets take advantage of the "convolution" operation on images, where the shift-invariant property, typical of the Euclidean domain is verified. Unfortunately, the shift-invariant property is hard to extend to non-Euclidean domains, where, a new *intrinsic* definition on a Riemannian manifold is necessary. Thus, recent works have tried to define a more general "convolution" for the non-Euclidean domain. Bruna et al. [42] proposed a spectral formulation of ConvNets on graphs. Masci et al. [139] proposed a generalization of ConvNets to triangular meshes using a local geodesic charting technique [149]. In this work, rather than defining a new "convolution" operation, we will use a traditional computer vision framework for 3D shape matching with the usual ConvNet architecture for RGB images.

In the traditional 2D setting we lose the view-invariant property. However, in cognitive neuroscience, Viewpoint-dependent theories suggest that object recognition is affected by the viewpoint at which it is seen, implying that objects seen in novel viewpoints reduce the accuracy and speed of object identification [281]. This theory of recognition suggests that objects are stored in memory with multiple viewpoints and angles. Unfortunately, the storage requirement increase as it requires that each viewpoint must be stored. The accuracy depends on how familiar the observed viewpoint of the object is [219]. Recent findings in the area of cognitive neuroscience in object recognition have established that the brain separates the object recognition process, from the face recognition [163].

Many methods have been proposed in computer vision to address the problem of limited

information in object recognition. Tremendous progress has been made especially in image-level object classification under limited geometric transformations, such as classification of side-view cars, or frontal view faces (e.g. [294],[92],[308]). Also relevant is the line of work in object detection in cluttered real-world scenes, such as pedestrian detection, or car detection [66], [91], [235], [302], [277], [53], [75]. However, most of the previous approaches can only handle up to a small degree of viewpoint variations of the 3D objects. As a result, they can hardly be used for robust pose understanding. A small, but growing, number of recent studies have begun to address the problem of object classification in a true multi-view setting [282],[302],[153],[298],[63],[257],[258],[172].

The proposed approach is based on a multi-view framework, where all the views are available in training and testing. Although inspired by the work of Su et al. [274], we are interested in evaluating the performances of a multi-views system for BFP classification. Differently from [274], we introduce a different rendering technique that jitters the data creating a more realistic and challenging environment.

A.2 Problem Definition

Given multiple RGB views of a given subject, our goal is to classify the subject concerning the BFP. Instead of an ordinary feature-based or shape-based representation, we will use the state-of-the-art in object classification: ConvNets. Feeding multiple images of the same subject to a ConvNet can make the network more prone to errors, resulting in a lower classification rate. In this work, we compare two modalities to feed the network. The first is view independent: we send to the network images of the same subject, from different viewpoints, independently, with no explicit correlation. This case is common to most classification problem, for instance like in ImageNet Challenge [250], CiFar-10 [152], Pascal VOC [87], etc. In our case, the

category is the BFP categories. The second model assumes some correlation between the views for each subject. The network knows that each subject has n views. This model has been implemented with a different network architecture, where an additional "pooling" layer works as views "accumulator".

A relevant part of this work is the generation of the needed views from the 3D dataset. In the next section, we describe this important step. Subsequently, we describe the network model and the training. Finally the results.

Pipeline

The tasks involved in this work are summarized in the following pipeline:

- Dataset:
  - **Rendering**: 16 views/subject with random illumination conditions.
  - **Organization**: split train/test 80%/20% set.
- Training:
  - **Setup pre-trained Network.**
  - **Fine-tune pre-trained Network (partial training).**
  - **Deploy the fine-tuned Network N. 1.**
  - **Fine-tune Network N.2.**
- Testing:
  - **Deploy Networks for testing.**
  - **Feature extraction on the test set.**
  - **Classification/Retrieval results.**

A.3 Dataset

We take advantage of the VirtualBody dataset introduced in Chapter 2. This dataset contains useful labels BFP, BMI, stature and body measurements. For the BFP classification, we use all the subjects in the Virtual NHANES dataset (Section 2.4.1) from all the weights classes. Differently, from the previous Chapter, we expect that the ConvNet will be invariant to scaling factors, such as the global dimension of the shapes. Figure A.2 shows some 3D male samples contained in the dataset. Figure A.3 shows 3D mesh and some of the 16 rendered views for one subject. In the next section, we explain the particular rendering process used to generate the views.

A.4 A Renderer for the VirtualBody Dataset

General 3D data is a representation of the content view-invariant. Today, this data format, with the excellent rendering capabilities of the GPU units, is giving outstanding performance in Virtual Reality (VR) [79]. Although, as humans, we cannot interact directly with 3D data. Moreover, to acquire and label the data is an expensive and very time consuming task, with the downside that some data can be scarce, noisy, and affected by some uncontrollable nuisances. In this context the computer vision community often resort to computer graphics techniques to solve more complex problems. One of the best example is the use of simulated human body pose depth images to train the random forest algorithm for body tracking in Microsoft Kinect SDK [266]. Figure A.1 shows the Microsoft Kinect depth map with body labels and skeletal joint points. Although typical random forest algorithms are not very demanding of training data, the problem requires the need of depth views of the body in a large variety of pose and activities. Such collection of data is quite demanding even for a corporate lab. The solution was to use virtual bodies, represented by meshes, and animate them using some MoCap data [67].

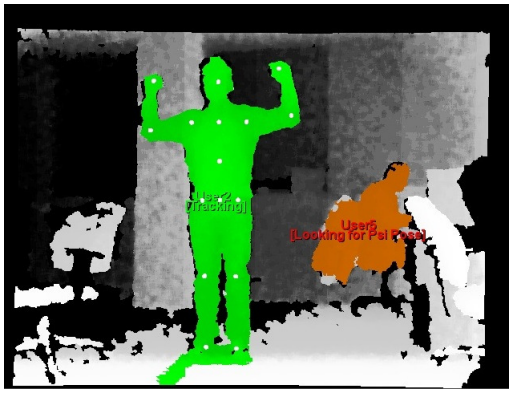


Figure A.1: Microsoft Kinect Body Tracking.

Using the virtual environment makes the body parts labeling accessible and more reliable than the expensive Vicon system. Inspired by this work we created VirtualBody, with the intention to simulate body shapes variation due mainly to BFP, and weight.

In Chapter 3 we propose an application based on computer graphics techniques to obtain many depth views of the subject. In this Chapter, we extend the method for the generation of thousands of views for the "hungry" (of training data) ConvNet algorithm.

A.4.1 Rendering

Rendering is the process of generating an image from a 2D or 3D model that we call subject. The subject is defined in a data structure (triangular mesh, for MH meshes). It would contain geometry, viewpoint, texture, lighting, and shading information as a description of the virtual scene. In our VirtualBody dataset, we define the textures for males and females, as for different races, but we do not define lighting, shading, and viewpoint. Then, we pass this information to a rendering program to be processed and output to a digital image or raster graphics image file. We define these quantities for the renderer: viewpoints, lighting and shading, and background.

The rendering process is similar to the raycast method introduced in Chapter 3. The main difference is that for the WBSA estimation we needed to produce 3D information as x, y, z co-

ordinates, then the raycast method is the preferable choice. We refer to the recent survey on rendering techniques [123]. The renderer is based on the pinhole camera model. Parameters of the model are the intrinsic parameters [114]. We do not use any distortion model in this work. An important parameter is the camera location. The position of the camera is defined as x, y, z position with respect to the origin, where the subject is located. We used, as in Section 3.2.2, a constant distance from the subject, and the angles: $0°, 30°, 45°, 60°, 90°, 120°, 135°, 150°, 180°, 210°, 225°, 240°, 27$ from the frontal position $(0°)$.

A.4.2 Data Augmentation and Jittering

Lighting and shading are two fundamental characteristics of the rendering process since they contribute to making the final result close to reality. Lighting or illumination is the artificial use of light to achieve a practical or aesthetic effect. Usually, the lighting is not part of the rendering equations, however many tools include some lighting model. The most famous is the Phong reflection model [220] (also called Phong illumination). It quantifies the surface reflected light as a combination of the diffuse reflection of rough surfaces with the specular reflection of shiny surfaces. Another critical element of the rendering process is the shading. Shading refers to the process of altering the color of an object/surface/polygon in the 3D scene, based on its angle to light sources and its distance from the light sources to create a photorealistic effect.

Lighting and shading, together, can make a drastic change on the appearance of the rendered subject. In our case, we use these two operations to jitter the data, creating a more realistic result. This data augmentation is very useful for training the ConvNets, which can easily "memorize" patterns in the data and overfit. Our solution is to randomly decide different illumination and shading conditions for each rendered view. Figure A.3 shows the original subject (center), and some rendered views. The views have different shading and illumination conditions, making the visual appearance almost unrecognizable from the original in some instances.

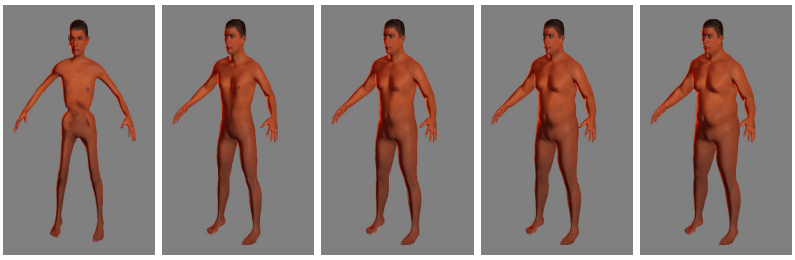


Figure A.2: 3D male models.

Background Overlaying: Till now we have generated multiple views of the same subject jittering the visual appearance with some lighting and shading artifacts. However, as we can see from Figure A.3, the silhouettes are still very clean because there is no background. This is a situation that can happen in reality. Thus the next step is to overlay a background image to the views. However, to avoid further "memorizing" effect by the ConvNet, for each rendered image, we randomly sample an image from SUN397 dataset [299] as the background image. We use the alpha-composition [228] to blend a rendered image as foreground and a scene image as background. Figure A.4 shows some of the results of the alpha-composition.

A.5 Network Models

One of the main goals of this Chapter is to compare the performance of two network architectures. The first one is the usual CNN network, and the second is a modified version of the first with a new pooling layer. We develop two experiments with the following specs:

- Experiment 1: Classification task (lean, fat, average weight) with a traditional fine-tuned CNN network on the independent views (CNN).

- Experiment 2: Classification task (lean, fat, average weight) multi-view CNN: the views are "pooled" at a pooling layer (MVCNN) as seen in [274].

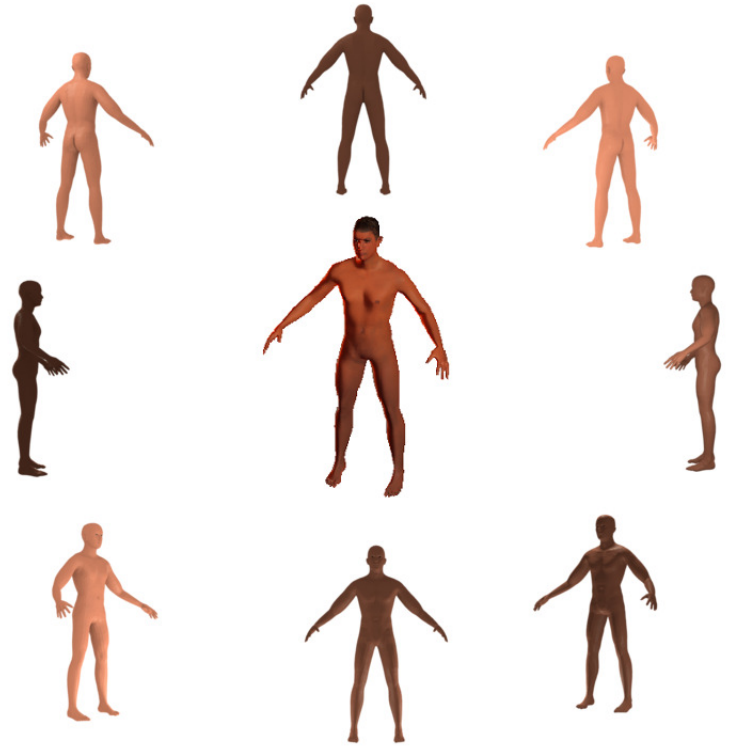


Figure A.3: Original mesh ans some of the 16 views.

Figure A.4: Background Overlaid for some of the views.

The second network differs in the multi-view pooling layer. A max-pooling kind of "pooling" able to **pool** together all the views before classification. The basic network used is the winner of Imagenet 2013 Zieler et al. [306], composed of 5 convolutional layers and 3 fully connected layers, with RELU between each layer and batch normalization (Figure A.5). This network, trained with Imagenet 2012, got the better results in the Imagenet 2013 competition.

The behavior of the two network for multi-view classification is different. A traditional ConvNet, trained with multiple views, will "accumulate" the information relative the view mostly in the convolutional layers, where a "collage" of diverse activation patterns constitute the activation function. Then, we can imagine different bodies from different views stored. The ConvNet with a multi-view pooling layer, instead, will have only the bodies information in the layers subsequent the pooling, since the pooling stage will aggregate all the view together.

A.5.1 Training

A.5.2 Data Partitioning

At the end of the rendering process, we obtained the new multi-view data collection. From this set we created two sets using an 80/20 % split between training and testing for both males and females experiments, obtaining the following:

- 12500 subjects, 16 views for every subject for a total of 200000 total images generated.
- 3 classes: Lean, Fat, Average (see Table A.1)

Unfortunately, given the number of images per class (16000) we cannot train this kind of network from random weights, because the data is not enough to train a network with this number of parameters ($\sim 90,000,000$). The number of parameters for this network is $> 90,000,000$

Class/Split	train	val	test
AVERAG	810 (16192)	202 (2000)	259 (2144)
FAT	810 (16192)	202 (2000)	259 (2144)
LEAN	810 (16192)	202 (2000)	259 (2144)

Table A.1: Number of subjects (images) per class.

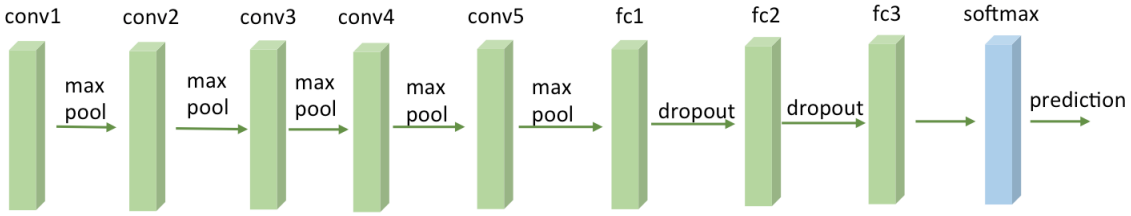


Figure A.5: ILSVRC 2013 Winner model.

the network will underperform due to the lack of training data. Thus the following choice is to use a pre-trained network and fine-tune only a few layers.

Fine-Tuning. The great power of deep ConvNets trained with a significative large amount of data is that they have a tremendous discriminative power and are invariant to very strong nuisances (e.g., affine transformations, illumination conditions, etc.). Unfortunately to train a network of this kind require a huge amount of data. A common solution, already in use with the old neural network architectures is to train only a portion of the network. This solution has a very strong motivation in the genesis of neural networks. In a large ConvNet, the initial layers are comparable to the primary elements of the visual cortex, V1 or striate cortex composed of simple cells. Recall that ConvNets and NNs, in general, are inspired by the visual cortex of mammals. Other areas are V2, V3, V4, V5, V6 or extrastriate areas, that account for much higher level processing. Image processing has taken advantage of these structures creating V1-based filters. Common V1-inspired filters (e.g., Gabor in image processing) are responsible for the detection of basic geometric elements: corners, edges, and scale-space analysis. In the

same fashion, the first few layers of a ConvNet work as a detector for basic geometric structures. With this in mind, it is possible to use an already trained network (e.g., for cats and dogs) on an entirely different dataset (e.g., pasta and cucumbers) just "tuning" some layers of the former network for the new dataset. Now the question is: "which layers that need to be trained again?". A common solution is to train the final layers of the network, and many use this policy to fine-tune the network with new data. See [126] for a review on fine-tuning techniques.

Batch Normalization. Training Deep Neural Networks is complicated by the fact that the distribution of each layers inputs changes during training, as the parameters of the previous layers change [132]. This slows down the training by requiring lower learning rates and careful parameter initialization, making notoriously hard to train models with saturating nonlinearities. This phenomenon has been called *internal covariates shift*. To address the problem, it is possible to normalize the inputs layers as Ioffe et al. suggested in [132]. Batch normalization helps in two ways: faster learning and higher overall accuracy. The improved method allows using a higher learning rate, potentially providing another boost in speed. The basis of this method is this intuition: we know that normalization (shifting inputs to zero-mean and unit variance) is often used as a pre-processing step to make the data comparable across features. As the data flows through a deep network, the weights and parameters adjust those values, sometimes making the data too big or too small (**internal covariate shift**). By normalizing the data in each mini-batch, the problem is mostly avoided. Basically, rather than just performing normalization once in the beginning, you are doing it all over the network.

A.5.3 Multi-View CNN: MVCNN

Inferring the shape of the subject from just one view can be a challenge when we have a single 2D RGB image. The principal problem in this situation is **missing information**. This problem

has been tackled in many ways. The most famous is matrix completion that in the specific 3D body setup has been largely discussed by the shape completion algorithm (SCAPE [7]). The same problem has been tackled in a different setup by the well known collaborative filtering techniques [311].

In this project, we follow the steps of the work by Su et al. [274]. The traditional ConvNet considers each view separately (many to many), and the scores from the views are averaged for the final result. The authors propose an attractive solution, where the views are pooled together for a unique decision. This solution introduces a new pooling layer, called "multi-view pooling", a max-pooling sort of layer. The novelty is quite intuitive. However, the paper misses a critical result. The new pooling layer has been tested on all the views, as it has been trained. It would have been way more interesting to see what happens if we use only one view during the test phase.

A.5.4 Testing: Deploying and Fine-tuning

Common terms used for the networks are deployed or trained. The training network is a network with the loss layer, which computes the error between the predicted and the actual output value. A deployed network, instead, does not have this layer, and the output labels are not propagated from the input of the network. The key tasks involved to fine-tune and test the networks (CNN,MVCNN) are the following:

- After fine-tuning, the network has to be deployed for testing (elimination of the layers used to train the network) (loss, argmax if SVM is used, etc..)
- To evaluate the classification performance, we trained an SVM classifier with the features extracted from the RELU7 layer.
- The MVCNN network, instead, is the deployed version of the former ConvNet with the

addition of the multi-views pooling layers.

Network Memory. Training a deep or very deep neural network is a task way more different than training a usual machine learning algorithm. It is something between, hacking, trial, and error, with some heuristic solutions. A good overview can be seen in the excellent tutorial by Bottou [34]. One of the most troublesome steps is the use of common tools (Caffe, Theano, Tensorflow, etc.) on the GPUs. The principal problem is the network footprint on the memory GPU. As we can see in Table A.2, with a huge amount of parameters, and many layers, the memory occupancy increases very quickly. The values in Table A.2 are relative to a batch size of 84 (num):

- Parameters Memory: 378MB (9.9e+07 parameters!)
- Data Memory: 1GB (for batch size 84)

As we can see, most of the memory is used by the batch of data that flow throughout the network. Unfortunately, the data cannot be removed from the memory after each layer since we still needed it to compute the derivative at the back propagation step! This is very important, and is one of the major constraints to implement very deep network on home computers. The parameters, luckily, do not take much of the memory since most of them are shared.

A.5.5 Results: Training

The above networks have been implemented using the matconvnet library [288]. The training time to fine-tune the network is around 6-7 hours for 15 epochs in the case of CNN, and around 3 hours to fine-tune the MVCNN network. Fine tuning the MVCNN network takes less time since only the last layer needs to be "tuned". We use a common I7 Desktop computer and an HPC machine with GPU GRID (4xK40) for heavy duty computing. In Figure A.6, we report the

Table A.2: Network Specs. M: Mbyte, K:Kilobyte, B:byte. Support define the convolutional layer geometry. The next rows: stride, size, and padding of the filters. Then the number, depth, and size of the filters. Finally the amount of data and parameters.

Layer	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
type	inp	conv	relu	mpool	conv	relu	mpool	conv	relu	conv	relu	conv	relu	mpool	conv	relu	conv	relu	conv	softm
name	n/a	conv1	relu1	pool1	conv2	relu2	pool2	conv3	relu3	conv4	relu4	conv5	relu5	pool5	fc6	relu6	fc7	relu7	fc8	loss
support	n/a	7	1	3	5	1	3	3	1	3	1	3	1	3	6	1	1	1	1	1
filt dim	n/a	3	n/a	n/a	96	n/a	n/a	256	n/a	512	n/a	512	n/a	n/a	512	n/a	4096	n/a	4096	n/a
num	n/a	96	n/a	n/a	256	n/a	n/a	512	n/a	512	n/a	512	n/a	n/a	4096	n/a	4096	n/a	55	n/a
stride	n/a	2	1	2	2	1	2	1	1	1	1	1	1	2	1	1	1	1	1	1
pad	n/a	0	0	0	1	0	0x1x0x1	1	0	1	0	1	0	0	0	0	0	0	0	0
rf size	n/a	7	7	11	27	27	43	75	75	107	107	139	139	171	331	331	331	331	331	331
rf off	n/a	4	4	6	10	10	18	18	18	18	18	18	18	34	114	114	114	114	114	114
rf stride	n/a	2	2	4	8	8	16	16	16	16	16	16	16	32	32	32	32	32	32	32
size	224	109	109	54	26	26	13	13	13	13	13	13	13	6	1	1	1	1	1	1
depth	3	96	96	96	256	256	256	512	512	512	512	512	512	512	4096	4096	4096	4096	55	1
num	84	84	84	84	84	84	84	84	84	84	84	84	84	84	84	84	84	84	84	1
data	48M	365M	365M	90M	55M	55M	14M	28M	28M	28M	28M	28M	28M	6M	1M	1M	1M	1M	18K	4
param	n/a	56KB	0B	0B	2MB	0B	0B	5MB	0B	9MB	0B	9MB	0B	0B	288MB	0B	64MB	0B	880KB	0B

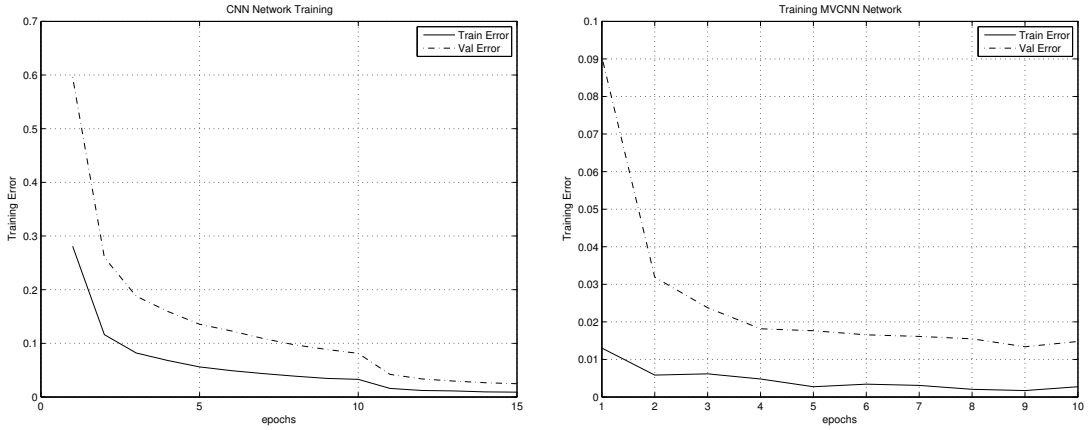


Figure A.6: Training and validation Errors. CNN (Left), MVCNN (Right)

training and validation error for the two networks. As we can see, the error dropped very fast with a few epochs. This behavior is characteristic of the fine-tuning procedure. A traditional network of the same dimension will take much longer to converge with randomly initialized parameters.

A.5.6 Results: Classification and Retrieval

In Table A.3 we report the results in testing for classification and retrieval task. In classification, the network is used to extract the features from the images. These features have been pulled from the network at RELU7 and used to train an SVM [33] classifier with a linear kernel.

The retrieval task is different. Based on information retrieval (IR), this requires a ranked list of items that are relevant to a specified query. The procedure is to compute the features distances between the query and all the other subjects, and subsequently rank them. Finally, common metrics are used to evaluate the performance. In this case, we report the mean average precision (mAP), and the area under the curve (AUC).

Table A.3: Classification and Retrieval results.

Results	Males		Females	
	CNN	MVCNN	CNN	MVCNN
Classification				
Accuracy (train)	99.95 %	100 %	99.49 %	100 %
Accuracy (test)	96.03 %	97.78 %	96.63 %	98.07 %
Retrieval				
mAP	88.89 %	91.54 %	90.62 %	91.22 %
AUC	88.08 %	91.20 %	90.59 %	91.19 %

A.5.7 Comparison with Spectral BFP

Table A.4 shows the classification results for the same weight grouping (W0-W1, all stature, all ages) used for the ConvNet approach. We can see that the ConvNet approach largely outperform the Spectral approach. This result is mainly due to the much more discriminative 2D "deep" features. However, the two methods are quite different. The spectral approach use only one 3D mesh, instead the ConvNet can leverage the more informative multi-view information.

Table A.4: Comparison with Spectral method.

Results	Males		Females	
	CNN	MVCNN	CNN	MVCNN
Classification				
Accuracy	96.03 %	97.78 %	96.63 %	98.07 %
Spectral BFP				
Accuracy	71.93 %		84.93 %	

A.6 Conclusion and Future Work

In this Chapter, we have presented an interesting application that use the recently proposed ConvNets to estimate the BFP class. We have introduced a rendering method to efficiently use the developed VirtualBody dataset in a common RGB framework. The newly rendered dataset allows us to develop endless multi-view methods since it can take advantage of the rich set of labels presented in Chapter 2. We have also tested two network configuration for the multi-view setting, obtaining excellent results. However, we found some important limitations in the original method. The networks are trained on all the subjects views, as most often happens in training time, but it is also tested on all the views. This last situation rarely happens in practice. Instead, a more interesting solution will be training on all the views and testing with only one view. One proposed work is to extend the present framework to test on a single random view.

Another interesting study is to consider the shape transformations as "style transfer". StyleNet [100] is one of nicest applications of ConvNets to the unusual field of artistic computer vision. Taking too picture, StyleNet can maintain the contest of the original picture, but transferring the style from the other picture. For a family of shapes, the style transfer can learn the "style" of fat people and transfer it to skinny people and vice versa. Boscaini [32] used the style transfer as a functional map for 3D shape retrieval but implemented on traditional spectral features.

Although the categorization in BFP classes is useful, a more interesting solution is the prediction of BFP. This measure, however, is hard to estimate accurately since it depends on many factors that the shape cannot account for (e.g., water in the body, the density of the bones, etc.). Instead, VirtualBody allows us to have an accurate measure of the WHR, which is officially recognized as an obesity indicator [9]. The proposed predictor, ideally will estimate the subject' WHR from a single view, after the training has been done with the multi-view approach.

Appendix B

Spectral Analysis

B.1 Helmotz Equation

$$\Delta f = \lambda f \tag{B.1}$$

The solutions of this equation represent the spatial part of the solutions of the wave equation. In the surface case $f(u, v)$ in Eq. B.1 can be understood as the natural vibration form (also eigenfunction) of a homogeneous membrane with the eigenvalue λ. The solutions of the general vibration problem are the solutions $f(u, v)$ of this differential equation on the surface. Because of this physical interpretation, the question whether the eigenvalues of the Laplace operator determine the shape of a planar domain, has been rephrased by the late mathematician L. Bers in a terse, impressively concise and pictorial way: Can one hear the shape of a drum? [229]
Δ is the Laplace-Beltrami Operator (LBO). Like the Laplacian, the LaplaceBeltrami operator is defined as the divergence of the gradient $\Delta f = \nabla^2 f = \nabla \times \nabla f$, and is a linear operator taking functions into functions. The operator is the generalization of the Laplace operator to Riemann manifold. From the spectrum of the Laplace-Beltrami operator, one can extract the area of S, the length of its border and its genus [247]. For a more deep understanding about the role of the

Laplacian operator on the manifold analysis is exciting the work of Canzani [48].

B.1.1 Spectrum Properties:

- The spectrum is defined to be the family of eigenvalues of the Helmholtz equation (Eq. B.1), consisting of a diverging sequence $0 \leq \lambda_1 \leq \lambda_2 \leq \lambda_3 \leq \ldots \inf$, with each eigenvalue repeated according to its multiplicity and with each associated finite dimensional eigenspace. In the case of a closed manifold without a boundary, the first eigenvalue λ_1 is always equal to zero, because in this case the constant functions are non-trivial solutions of the Helmholtz equation. If a Dirichlet boundary exists, the first eigenvalue is always greater than zero, since the only constant solution is trivial (because of the boundary condition). The first eigenvalue is always simple, and the corresponding eigenfunction has no nodal lines (zero sets of the function). The nodal lines of the nth eigenfunction subdivide the domain into maximal n subdomains.

- The spectrum is an isometric invariant as it only depends on the gradient and divergence which in turn are defined to be dependent only on the Riemannian structure of the manifold. This implies property **ISOMETRY**.

- Furthermore, we know that scaling an n-dimensional manifold by the factor a results in scaled eigenvalues by the factor $1/a^2$. Therefore, by normalizing the eigenvalues, the shape can be compared regardless of the objects scale (property **SCALING**). This fact can be proved quite easily for any dimension n.

  Let M be a compact n-dimensional Riemannian manifold of class C^N with the local parametrization $h : \mathbb{R}^n \rightarrow \mathbb{R}^{n+k}$. The scaled manifold with the parametrization $\bar{h} : ah$

possesses the partial derivatives

$$\partial_k \bar{h} = a\partial_k h (k = 1, \ldots, n) \text{ implying } g^{-ij} = \frac{1}{a^2} g^{ij} \text{ and } \bar{W} = a^2 W, \tag{B.2}$$

$$\Delta_h u = \frac{1}{W} \sum_{i,j} \partial_i (g^{ij} W \partial_j u) = \lambda u \tag{B.3}$$

$$\Delta_{\bar{h}} u = \frac{1}{\bar{W}} \sum_{i,j} \partial_i (g^{ij} \bar{W} \partial_j u) = \frac{1}{a^2 W} \sum_{i,j} \partial_i (g^{ij} W \partial_j u) = -\frac{1}{a^2} \lambda u \tag{B.4}$$

- The spectrum depends continuously on the shape of the membrane, thus complying with property **SIMILARITY**. Moreover, it can be shown with similar arguments that the spectrum depends continuously on the **Riemannian metric** of the manifold in general.

- The numerical computation of the spectrum can already be done with a standard personal computer. Therefore the requested **EFFICIENCY** can be satisfied as well.

- The property **COMPLETENESS** is not fulfilled by the spectrum because some non-isometric manifolds with the same spectrum exist.

- The question if a sequence of n real numbers ($S = \{a_1 = 0 \leq a_2 \leq a_3 \leq \cdots \leq a_n\}$) can be the beginning of the spectrum of a compact Riemannian manifold X has been discussed by Colin de Verdière. It is shown that for any such finite sequence S, there always exists a compact Riemannian manifold X with $dim(X) \geq 3$ always exists realizing S as the beginning of its Laplace spectrum. This result also means that given any positive integer n, a Riemannian manifold exists, such that the multiplicity of the first non-zero eigenvalue is n. In the case of a closed Riemannian surface ($dim(X) \geq 2$), there are bounds to the multiplicities depending linearly on the genus. However, in the

case of a surface, the result by Colin de Verdière holds also for finite sequences of the form ($S = \{a_1 = 0 \leq a_2 \leq a_3 \leq \cdots \leq a_n\}$). These results are interesting in the context of property **COMPRESSION**.

Of course, classes of manifolds exist (like the disks or the rectangle) where one or two eigenvalues already determine the size and the shape and therefore the whole spectrum. In other words, if we know we have a rectangle, we need just two eigenvalues to find its side lengths. *Without prior knowledge of the manifold, a characterization is impossible by a finite subsequence of the spectrum*. Therefore, the spectrum cannot be compressed into a finite subsequence (see property **COMPRESSION**) without losing information.

- A substantial amount of geometrical and topological information is known to be contained in the spectrum. Therefore the property **PHYSICALITY** is fulfilled. Even though we cannot crop a spectrum without losing information, we will show that it is possible to extract important information just from the first few eigenvalues (approx. 500).

However, it will not be possible to satisfy property [COMPLETENESS]. Nevertheless, no three pairwise isospectral but non-isometric manifolds have been constructed so far and all known pairs of isospectral planar domains have been shown to be non-convex with non-smooth boundaries. The only examples of pairs of convex domains in Euclidean space, being isospectral but not congruent, were found in four or higher dimensional spaces. It is not sure if triples or isospectral continuous deformations exist in lower dimensions at all. The constructed examples (e.g., pairs of isospectral domains) were always somewhat artificial and appear to be exceptional. For the special case of Riemann surfaces (namely surfaces with constant negative curvature), Buser was able to derive an upper bound for the number of isospectral but non-isometric surfaces depending only on the genus. For all of these reasons and also based on experimental studies, we feel that the spectra of the LaplaceBeltrami operator have significant

discrimination power, strong enough to be used in contemporary applications, as point clouds. This will be the most recurrent data in future surveillance systems, acquired from mobile or fixed lidar devices, RGB-D sensors, or simply as multi-view stereo reconstruction.