COMPUTATIONAL APPEARANCE MODELS FOR QUANTITATIVE DERMATOLOGY

BY PARNEET KAUR

A dissertation submitted to the

Graduate School—New Brunswick

Rutgers, The State University of New Jersey

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

Graduate Program in Electrical and Computer Engineering

Written under the direction of

Prof. Kristin Dana

and approved by

_____

_____

_____

_____

_____

New Brunswick, New Jersey

October, 2017

ABSTRACT OF THE DISSERTATION

Computational Appearance Models for Quantitative Dermatology

by Parneet Kaur

Dissertation Director: Prof. Kristin Dana

Skin appearance modeling using high-resolution imaging has led to advances in recognition, rendering and analysis. In dermatology, workforce shortage and long patient wait time has motivated the need for computational methods to assist the dermatologists. In recent automated image recognition tasks, deep learning with convolutional neural nets (CNN) has achieved remarkable results. However in many clinical settings, training data is often limited and insufficient for CNN training. Furthermore, skin images have subtle differences and are very different from the typical images used for computer vision tasks. This motivates the need of developing methods that can be used for limited and unique datasets. In this research, we propose computational models using deep learning approaches for novel problems in quantitative dermatology.

First, we develop a photo-realistic facial style transfer method (FaceTex), which transfers facial texture from a new style image while preserving most of the original facial

structure and identity. FaceTex has implications in commercial applications and dermatology, such as visualizing the effects of age, sun exposure, or skin treatments (e.g. anti-aging, acne). We suppress the changes around the meso-structures (eyes, eyebrow, nose, lips and lower facial contour) by introducing the Facial Prior Regularization that smoothly slows down the updating. Additionally, we tackle the challenge of preserving facial shape by minimizing a Facial Structure Loss, which we define as an identity loss from a pre-trained face recognition network that implicitly preserves the facial structure. Our results demonstrate superior texture transfer than state-of-the-art methods because of the ability to maintain the identity of the original face image.

Second, we develop a computational skin texture model to characterize image-based patterns from ultraviolet and blue fluorescence multimodal images and link them to distribution of microbes on the skin surface, i.e. the skin microbiome. The intersection of appearance and microbiome clusters reveals a pattern of microbiome that is predictable with high accuracy based on skin appearance. We present a new approach, appearance-driven multiview co-clustering (AMCO), which incorporates both multiview and co-clustering in order to discover which microbiome parameters are linked to appearance.

Finally, to measure the thickness of skin layers, we develop a hybrid deep learning method to classify reflectance confocal microscopy images. We also use CNNs to classify the images and demonstrate that smaller training datasets are insufficient for CNN training and feature extraction is essential in such cases. We compare our method with a suite of texture recognition methods for RCM images and show that hybrid deep learning outperforms the state-of-the-art with a test accuracy of 81.73%. Using a patch-based approach and pre-trained CNNs for feature extraction, we achieve a peak classification accuracy of 89.87%.

Acknowledgements

Along the journey of this dissertation, I have been helped, encouraged and inspired by several people. Here, I want to thank everyone who contributed to its development and completion.

First and foremost, I owe my deepest gratitude to my advisor Prof. Kristin Dana. She believed in my capabilities and gave me the opportunity to purse my PhD degree. She has made available her support, time and wisdom in a number of ways throughout this research. I thank her for her consistent guidance and constructive feedback.

I would also like to express my gratitude to the rest of my thesis committee: Prof. Emina Soljanin, Prof. Saman Zonouz, Dr. Gabriela Oana Cula and Dr. Michael Isnardi, for their insightful questions and comments.

I gratefully acknowledge the financial support received from Johnson and Johnson Consumer Products Research and Development. I am thankful to the collaborators from Johnson and Johnson Consumer Products Research and Development. I would especially like to thank Dr. Gabriela Oana Cula for numerous discussions and valuable feedback. I also express my gratitude to Dr. Dianne Rossetti, Dr. Kimberly Capone and Dr. Nikoleta Batchvarova, for their contribution towards clinical design, data acquisition and many valuable conversations about the microbiology aspect of the data. Special thanks to Dr. Catherine Mack for her contributions to the reflectance confocal microscopy dataset.

I am very thankful to Google and WomenTechMakers for supporting me with the Google Anita Borg Memorial Scholarship. I am very honored to be a recipient of this scholarship

and it had a very positive impact on my personal and professional development.

I am grateful to the department of Electrical and Computer Engineering (ECE) for financially supporting me as a teaching assistant and a fellow. I thank the past and present administrative staff members of the ECE department for their prompt support and kindness: John Scafidi, Noraida Martinez, Arletta Hoscilowicz, Christy Lafferty, Tea Akins, Ora Titus and Steve Orbine.

I am thankful to members of the computer vision lab - Hang Zhang, Eric Wengrowski, Jia Xue, Thomas Shyr, Matthew Purri and Hansi Liu, for providing a comfortable work environment and valuable discussions.

I thank Johnson and Johnson Consumer Products Research and Development for providing multimodal skin images, skin microbiome and reflectance confocal microscopy dataset. I also thank photographer Martin Scheoller and Art+Commerce for allowing us to use their photographs for texture transfer.

Finally, I would like to thank my family and friends. I want to thank my parents and brother their unconditional love and encouragement. Even though they are thousands of miles away, they were always here whenever I needed them. My parents-in-law have been very patient and prayed for my success. My dear husband, Jaspreet, has been my best friend and a great companion. This thesis would not have been possible without his love, encouragement and help. My daughter, Gurbani, has patiently shared my time and attention. She has given me daily dose of motivation with her beautiful smile and lots of hugs. I also thank my friends, Sudha, Manisha, Parul, Sumati and Anvita, for providing all the moral support and friendship that I needed. Last but not the least, I thank my VaddeMama (paternal grandmother) and NaniMa (maternal grandmother) for their blessings.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Skin is the largest organ of the human body. Its multi-layered structure provides a protective covering to the body, shielding internal body organs and tissues from trauma, ultraviolet radiation (sun exposure) and harmful bacteria. It also prevents dehydration, regulates body temperature and synthesizes vitamin-D. Like any other body organ, skin is prone to disorders and diseases, and requires medical as well as preventive care. To assist dermatologists and make the skin analysis robust and faster, computational methods using computer vision and machine learning have been proposed in literature. However, recent methods of deep learning, which have led to a significant improvements in several computer vision applications have not been fully explored in the field of dermatology.

In this dissertation, we propose deep learning approaches for three novel problems in dermatology. First, we use convolutional neural networks based approach for texture transfer between facial images. This method can be used to visualize the effects of aging, sun exposure or skin treatments. Second, we develop computational skin appearance models from images captured in different imaging modalities and link them to a biological measurement, skin microbiome, i.e., the distribution of bacteria on our skin. Finally, we propose hybrid deep learning and patch-based convolutional neural networks to measure the thickness of skin layers.

In this chapter, we present the background and motivation of our research by discussing the necessity of advanced computational methods in quantitative dermatology. The rest of this chapter is organized as follows: Section 1.1 explains the structure and functionality of different skin layers. Section 1.2 provides an overview of the skin microbiome, i.e., the bacteria present on the surface of our skin. Section 1.3 discusses skin conditions and diseases, and how changes in skin appearance can be an indication of internal diseases. In

Section 1.4, we discuss the state of skin healthcare and need to assist the dermatologists for diagnoses and better understanding of skin using quantitative methods. Section 1.5 presents a brief overview of deep learning and in particular, convolutional neural networks. Finally, Section 1.6 provides an overview of how we leverage and extend the deep learning methods for quantitative dermatology in this research.

1.1 Skin Anatomy

Human skin consists of three layers, each performing specific functions [139]. The thickness of each layer varies for each individual and body parts. The three layers are:

Epidermis: It is the topmost layer of the skin, which provides a protective barrier from the external environment. It mostly consists of keratinocytes cells, which migrate from the lower layers of epidermis towards the upper layers and gradually shed. They prevent loss of moisture, regulate temperature and shield body from pathogens. Melanocytes cells in epidermis contain melanin, which is also responsible for the skin color. Darker skin has more amount of melanin than lighter skin. More importantly, melanin filters out ultraviolet radiation from sunlight and prevents DNA damage. Langerhans cells are antigen-presenting cells, which are activated during skin infections or cell damage. Merkel cells in lower layers of epidermis are responsible for sensation of light touch.

Dermis: The uppermost layer of dermis, papillary dermis, is mainly composed of elastic fibers whereas the lower layer, reticular dermis, contains collagen fibers. Both collagen and elastic fibers are produced by fibroblast cells and give the skin strength and elasticity. Dermis also contains nerves, hair follicles, sweat glands, sebaceous glands and blood vessels. These dermis components vary for different body parts. For example, scalp may have a lot of hair follicles but palm of hand may have none. Nerve endings in dermis respond to stimuli. The sebaceous glands secrete sebum (oil), which lubricates the skin. The sweat gland can either be apocrine glands or eccrine glands. The eccrine glands help to regulate body temperature when it rises, by secreting a fluid on skin surface that cools the body as it evaporates. Apocrine glands secrete a milky fluid which gets combined with the bacteria found on the skin and produce body odor.

Hypodermis (Subcutaneous Layer): It consists of fatty tissue, which protect the body from extreme temperature and cushions internal body organs. It also consists of blood vessels and nerves.

1.2 Skin Microbiome

Human skin hosts a variety of bacteria, viruses, fungi, archaea and small arthropods [49, 96, 114]. Using 16S rRNA gene sequencing, recent research has established the role of skin microbiota with diseases as well as its significance to maintain a healthy skin [48, 65, 69, 96]. Studies of the skin microbiome show dependence on genetics, environment and behavior as well as a variation over time [15, 66, 68, 142, 156]. The skin microbiome varies for each individual and also for location on the body [66]. The studies of infant skin [13] show a distinct difference in infant microbiome as a function of age, suggesting an evolution of the microbiome in the early years of development. Contrarily, adult microbiome is generally stable over time and establishing a healthy stable skin microbiome assists in resisting transient harmful microbes. A symbiosis of the human immune system and the skin microbiota has been explored [81].

Interactions between host and microbes as well as microbe-to-microbe interactions have an impact on skin disorders and diseases such as psoriasis, acne, dermatitis, and rosacea [142, 198]. *Propionibacterium acnes*(*P. acnes*) has long been associated with the presence of acne. Recent studies of human skin microbiome show that some strains of *P. acnes* are associated with acne while others are present in abundance in non-acne skin [48]. Another skin disorder, atopic dermatitis (eczema) has been associated with specific bacterial species called *Staphylococcus aureus*. By analyzing the skin microbiome obtained during the treatment of atopic dermatitis patients, topographic and temporal shift in the predominant bacterial species and the bacterial diversity have been demonstrated [96].

With the advent of fast and cost-effective DNA sequencing technology, modern methods for characterizing the skin microbiome are genomics-based where skin swabs are sequenced for detection of microbial DNA [50, 180]. Currently, swabbing and sequencing is cost-prohibitive for routine imaging. Prior methods use culture-based characterization of

microbes that are biased toward "weeds", the microbes that grow quickly and easily in isolation [96]. However, while many microbes can thrive within a microbial community, only a few species grow well in isolation.

Study of skin microbiome and its association with specific skin disorders raise possibilities for entirely new directions of skin treatments [65]. Molecular-based methods are ushering in an age of metagenomics [133,134], where microbial DNA is quantifiable and can be used for analysis or for engineering therapies. Potential opportunities in drug discovery can be based on the concept that inflammatory skin disease may be associated with the disturbance of the skin microbial equilibrium and methods can be engineered to modify or stabilize the microbiome [163]. Further research in skin microbiome and study of useful and harmful microorganisms can pave pathway for precision and personalized medicine.

1.3 Skin disorders, diseases and skin manifestation of internal diseases

Skin disorders can be caused by a variety of factors including irritation or clogging of skin, inflammations, infections and internal diseases. Skin problems or diseases typically result in lesions of different shapes, sizes, colors and severity on varying skin locations [80]. Acne is one of the most skin disorder caused by plugging of the hair follicles with oil secreted by the sebum glands and dead skin cells. The clogged pore becomes infected with bacteria and results in inflammation. Another inflammatory skin problem, atopic dermatitis (Eczema), results in dry and itchy skin. Some skin diseases effect certain age groups. For example, children are susceptible to viral and bacterial skin infections like measles and impetigo. Viral infection like shingles can be seen in adults, which is caused by the relapsing of the virus that causes chickenpox. For severe diseases include melanoma and non-melanoma (basal and squamous cell carcinoma) skin cancers. These skin cancers can be caused by chronic exposure to sun (ultraviolet light), ionizing radiation and chemical carcinogens such as arsenic and tobacco [36]. While some of the skin disorders and diseases can be easily treated or controlled, diseases like skin cancer may require surgery, chemotherapy or radiation therapy.

The problems and changes in the skin are not limited to skin disorders and may be

an indication of underlying internal diseases [10, 80]. For example, any new lesions on the skin can indicate spread of internal cancer or other diseases to the skin [59]. Addison's disease causes hyperpigmentation and is caused by insufficient production of cortisol and aldosterone hormones [61]. Yellowing of the skin can indicate liver disease. Bronzing of the skin in a patient with diabetes can be a sign of hereditary hemochromatosis, where defect in iron metabolism can lead to liver or heart failure [6]. An unusual change in the texture or thickness of the skin can also indicate an internal problem. Swelling and then hardening of the skin are early signs of systemic sclerosis- an autoimmune disease [155]. On the other hand, imbalance in thyroid hormone results in thin skin [160].

Diagnosis of the skin disorders and diseases requires careful inspection by the dermatologists. Early diagnosis, treatment as well as prevention can improve prognosis and decrease the effect of these disorders on the patients.

1.4 State of Skin Healthcare and Need for Quantitative Dermatology

The Global Burden of Disease Study in 2013 identified skin diseases as one of the significant contributor of non-fatal global disease burden [87]. In 2013, the direct health care cost of skin conditions and diseases was calculated as $75 million [117]. This direct cost includes the amount paid by the insurance plan and the amount paid by the patient for deductibles, coinsurance and copayments. This amount was used for diagnosis, evaluation and management services, skin procedures, skin disease vaccination, prescription medicines and over the counter products. Additionally, the indirect costs (loss of wages) caused by interactions with the healthcare were estimated as $11 billion.

Skin conditions and diseases require careful inspection, analysis, evaluation and diagnosis by dermatologists. Several studies have reported a shortage of dermatologist workforce resulting in a long wait time for the patients [18,93,153,154]. In United States, new patients had to wait for 33 calendar days on an average for an appointment with a dermatologist in 2007 [93]. The average wait time for dermatologist appointment has only marginally decreased by 3 days since 2002 [153]. Similar undersupply has been reported for pediatric

dermatologists in academic as well as clinical settings [18]. Typically, 71% of the dermatologists spend most of their time on medical dermatology whereas 29% spend most of their time on surgery and cosmetic dermatology [93]. Furthermore, in clinics and hospitals that do not provide dermatology services, teledermatology can be used to make skin care accessible [4]. Clearly, there is a need to address the shortage of dermatology experts.

In addition to increasing the pool of dermatologists, several innovative alternatives such as teledermatology, point-of-care diagnostic tools and task shifting have been suggested in [165]. In addition to these alternatives, assisting the dermatologists using automated quantitative evaluation can help to reduce per patient care time. Many dermatologists manually analyze skin conditions and diseases from skin images, which is time consuming. Moreover, there can be interobserver and intraobserver differences in diagnoses of same skin image [77]. Adding a quantitative measure using computational analysis can assist the dermatologists in decision making. Modern computer vision and machine learning methods can be used for computer-assisted diagnosis of skin images to ensure that needs of patients with skin conditions and diseases are met in a timely manner. These methods can also be used to get a better understanding of skin from images for preventive dermatology [143]. Finally, visual skin analysis can be used as an interface between patients and dermatologists to better explain the effect of using a certain skin treatment or product.

1.5 Deep Learning in Computer Vision

In artificial intelligence, the fields of computer vision and machine learning (Figure 1.1) have been integrated to understand and interpret the real world scenes with applications such as object recognition, tracking, visual surveillance and scene reconstruction. In computer vision, images are acquired using sensors such as camera, radar, etc. and then computationally analyzed to sense and understand the surroundings. In machine learning, a set of rules are learned from the data using computational algorithms as opposed to traditional programming where set of rules are used to process data faster.

In traditional computer vision, images are processed to represent an image as informative feature descriptors (Figure 1.2). Many approaches use local image gradient information,

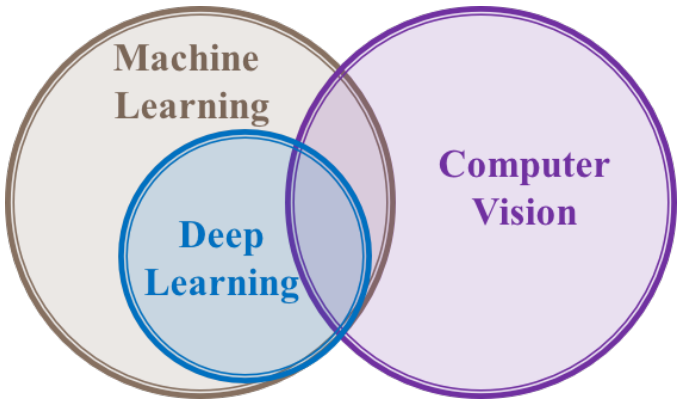


Figure 1.1: **Computer Vision, Machine Learning and Deep Learning.**

which extract low-level features such as SIFT [124] and HOG [29]. More structured and global information is captured by the mid-level features such as bag of features [173] and spatial pyramids [107]. High-level features utilize image semantics and image content [116, 158]. These feature representations are often referred to as hand-crafted features and are selected to solve a specific problem based on intuitions for the datasets. For an image classification task, a traditional classifier such as support vector machines can be trained using these feature vectors to distinguish between different categories.

Recently, deep learning and in particular, convolutional neural networks have advanced the performance of several tasks in computer vision like image classification and recognition [75, 100, 171, 179], object detection [123] and image segmentation [3]. Deep learning uses several hierarchical layers to learn useful feature representation of data.

Artificial Neural Networks (ANNs) are inspired by biological neurons and were introduced in 1950s. CNNs are an extension of on ANNs for images and were first introduced for handwritten character recognition in 1998 [108]. Since they required a lot of training data and training time, they could not be expanded to other computer vision tasks at that time. After a hiatus of over a decade, CNNs resurfaced in 2012 [100] due to advancements in technology. First, more data was made available to the research community. Imagenet [157] was released in 2009 for an image classification challenge, which consists of 1000 object classes, 1.2 million training and 100,000 test images. Secondly, graphical processing units (GPUs) provided a lot more computing power. Finally, the availability of more data, increased computing power and promising results resulted in more involvement of research community,

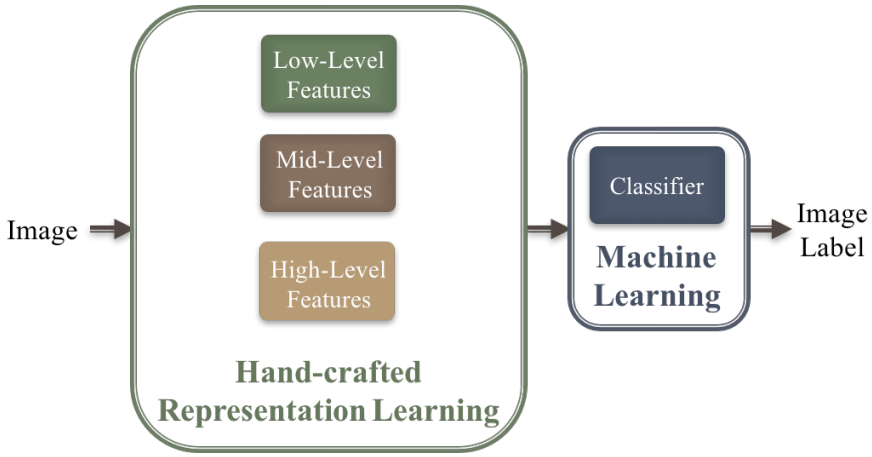


Figure 1.2: **Traditional computer vision pipeline for image classification.** Hand-crafted features are extracted to represent images and used for training a typical machine learning classifier like support vector machines.

which led to breakthroughs in the field of deep learning. Figure 1.3 shows the top-5 error rate for the test classes of the Imagenet dataset [75, 100, 171, 179]. Top-5 error is the percentage of test images whose target (true) label does not match any of the top-5 predicted labels. In 2010 and 2011, shallow networks with traditional hand-crafted features were used and resulted in 28.2% and 25.8% top-5 error on Imagenet test set. In 2012, AlexNet [100] was the first deep network used on ImageNet, which dropped the error to 16.4%(9.4% less than in 2011). Since then several deeper networks [75, 171, 179] have been proposed and improved the performance. At the same time, the network depth i.e., the number of layers in the network have increased from 8 in 2012 to 152 in 2015. The top-5 error on Imagenet test data also surpassed human performance in 2015.

CNNs are typically used for image classification tasks and trained on a huge training dataset such as ImageNet. The dataset is labeled i.e., each image has a class label associated with it. Once the network is trained, the weights of each layer are stored and used during the forward pass to predict the class of test images. A trained CNN, whose network weights have been saved, is referred to as a *pre-trained network*. Since the weights learned using huge datasets such as ImageNet generalize well, pre-trained CNNs can also be used for transfer learning. If a new dataset consists of lot of training data, we can initialize the CNN weights with the pre-trained network (trained on ImageNet) and then re-train the network. Initializing the weights from pre-trained networks results in faster convergence. If the new

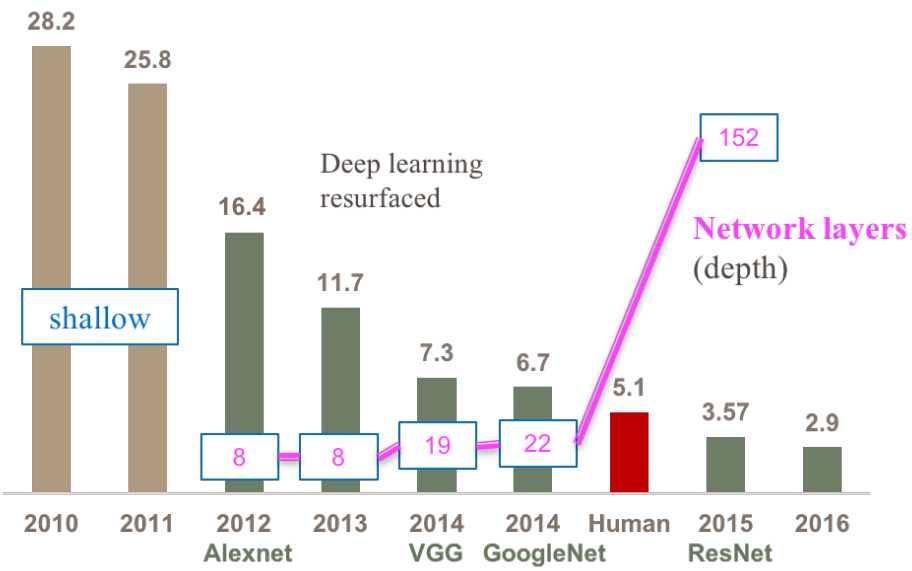


Figure 1.3: **Top-5% error for ImageNet.** Deep learning network dropped the error by 9.4% in 2012 [100]. Since then several deeper networks [75, 171, 179] have been proposed and improved the performance. The network depth i.e., the number of layers in the network have increased from 8 in 2012 to 152 in 2015.

dataset is small, CNNs can be used to extract feature vectors to train a traditional classifier.

A typical CNN architecture consists of multiple stacks of convolutional layers, rectilinear linear unit and pooling layer, followed by one or more fully connected layers as shown in Figure 1.4. The typical layers are:

Convolutional Layer consists of multiple learnable filters (kernels) with a small receptive field, which extends to the depth of the input. The filter size decides the area of the receptive field. The entire input is convolved with each of these filters to obtain an activation map. All activation maps are stacked together as output.

Pooling Layer is used to reduce dimensionality of the feature maps in each layer by downsampling the data. Some of the commonly use pooling algorithms are max pooling and average pooling pooling. Max pooling extracts the maximum value while average pooling takes the mean of all values in each subregion of the feature map.

Rectifier Linear Unit (ReLU) Layer introduces non-linearity in the network by thresholding all the negative values of the gradients to zero when the unit is not active. It leads to a faster convergence during training [100] as compared to sigmoid activation. Leaky ReLU

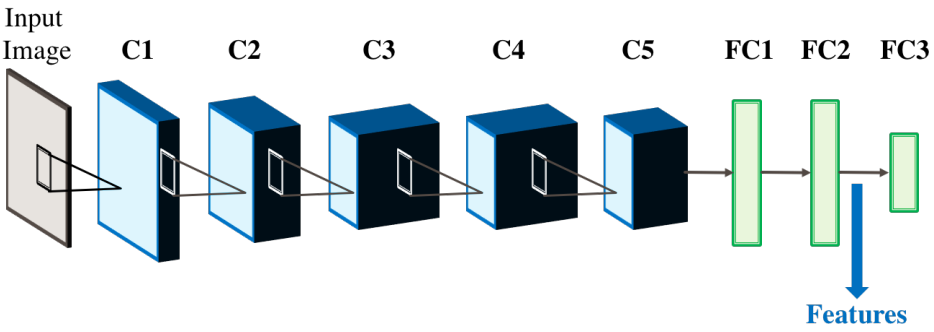


Figure 1.4: **Convolutional Neural Networks.** A typical CNN architecture consists of multiple stacks of convolutional layers, rectilinear linear unit and pooling layer, followed by one or more fully connected layers. Here, C1...C5 represent convolutional, maxpool and ReLU layers. FC1, FC2 and FC3 are fully connected layers.

allows a small, non-zero gradient instead of making it zero [126]. In Parametric Rectified Linear Unit (PReLU), the slope in the negative region is parameterized for each neuron and learnt with rest of the network [74].

Fully Connected Layer allows each node to be connected to all the nodes in the preceding layer. Since this is a dense layer with a lot of node connections, it increases the parameters. The final layer has same number of nodes as the number of classes in the training dataset. This layer typically has softmax loss for predicting a single class for mutually exclusive classes or sigmoid cross-entropy loss for predicting independent probability values of each class.

1.6 Thesis Overview

The computational methods in computer vision and machine learning that have led to advances in quantitative dermatology. However, the recent deep learning methods have not been fully explored for automated skin image analysis. Developing deep learning methods usually require a lot of training data, which is not usually available for dermatology studies. While datasets such as ImageNet can be labeled using online tools (for example, Mechanical Turks), labeling of skin images requires experts with prior knowledge. The fine-grained, subtle differences in the skin textures are not common in typical computer vision datasets. Hence, careful insights are required while developing deep learning methods with small and

unique datasets. In this research, we propose computational models using deep learning approaches for three novel problems in quantitative dermatology.

Texture Transfer for Facial Images: Style transfer methods have achieved significant success in recent years with the use of convolutional neural networks. However, many of these methods concentrate on artistic style transfer with few constraints on the output image appearance. We address the challenging problem of transferring face texture from a style face image to a content face image in a photorealistic manner without changing the identity of the original content image. This method can be used to visualize the effects of aging, sun exposure or skin treatments. Our framework for face texture transfer (FaceTex) augments the prior work of MRF-CNN with a novel facial semantic regularization that incorporates a *face prior regularization* smoothly suppressing the changes around facial meso-structures (e.g eyes, nose and mouth) and a *facial structure loss function* which implicitly preserves the facial structure so that face texture can be transferred without changing the original identity. We demonstrate results on face images and compare our approach with recent state-of-the-art methods. We evaluate our results qualitatively as well as using two quantitative measures: landmark error and texture similarity. Our results in Chapter 2 demonstrate superior texture transfer because of the ability to maintain the identity of the original face image. These results have been presented in [90].

Skin Appearance and Skin Microbiome: Skin appearance modeling using high resolution photography has led to advances in recognition, rendering and analysis. Computational appearance provides an exciting new opportunity for integrating macroscopic imaging and microscopic biology. Recent studies indicate that skin appearance is dependent on the unseen distribution of microbes on the skin surface, i.e. the skin microbiome. There has been interest in research community to gain better understanding of skin microbiome to use it for precision and personalized medicine. While modern sequencing methods can be used to identify microbes, these methods are costly and time-consuming. We develop a computational skin texture model to characterize image-based patterns and link them to underlying microbiome clusters. The pattern analysis uses ultraviolet and blue fluorescence multimodal skin photography. The intersection of appearance and microbiome clusters reveals a pattern of microbiome that is predictable with high accuracy based on skin appearance.

Furthermore, the use of non-negative matrix factorization allows a representation of the microbiome eigenvector as a physically plausible positive distribution of bacterial components. Chapter 3 presents the first results in this area of predicting microbiome clusters based on computational skin texture. These results have been published in [92].

After establishing the link between skin appearance and microbiome, we address the problem of causative appearance analysis. We address the problem of linking visual appearance to a secondary or auxiliary set of parameters obtained from a different source. The underlying and novel goal is to go beyond appearance labeling to discover associations that reveal how parameters affect appearance. We develop a new approach that leverages and extends methods of multiview clustering and co-clustering. In multiview clustering, multiple views of a subject from multiple sources are clustered simultaneously and agreement is made between cluster membership in the spaces. Co-clustering has a different goal, the discovery of which features are used to create groups, but in a single space. We present a new approach, appearance-driven multiview co-clustering (AMCO), which incorporates both multiview and co-clustering in order to discover which secondary space parameters are linked to appearance. Results are demonstrated on appearance-microbiome and animals with attributes datasets in Chapter 3.

Classification of Microscopic Skin Images: Reflectance Confocal Microscopy (RCM) is used for evaluation of human skin disorders and the effects of skin treatments by imaging the skin layers at different depths. Traditionally, clinical experts manually categorize the images captured into different skin layers. This time-consuming labeling task impedes the convenient analysis of skin image datasets. In recent automated image recognition tasks, deep learning with convolutional neural nets (CNN) has achieved remarkable results. However in many clinical settings, training data is often limited and insufficient for CNN training. For recognition of RCM skin images, we demonstrate that a CNN trained on a moderate size dataset leads to low accuracy. We introduce a hybrid deep learning approach which uses traditional texton-based feature vectors as input to train a deep neural network. This hybrid method uses fixed filters in the input layer instead of tuned filters, yet superior performance is achieved. Our dataset consists of 1500 images from 15 RCM stacks

belonging to six different categories of skin layers. We show that our hybrid deep learning approach performs with a test accuracy of 82% compared with 51% for CNN trained from scratch. It is demonstrated that smaller training datasets are insufficient for CNN training and feature extraction is essential in such cases. These results have been published in [91]. Further, using pre-trained CNNs and a patch-based approach, accuracy of 85.53% is achieved. Furthermore, we highlight the ambiguity in labeling at transition regions and achieve 89.87% accuracy by allowing mislabeling at the transition regions. We also compare the results with additional proposed methods for RCM image recognition and show improved accuracy. These results are presented in Chapter 4.

Finally, Chapter 5 discusses the conclusions and future direction of this research.

Chapter 2

Texture Transfer for Facial Images

2.1 Introduction

Recent work in texture synthesis and style transfer has achieved great success using convolutional neural networks [54, 55]. In style transfer techniques, the style-transformed image is synthesized to maintains contents of the original image while transferring style from another image. Figure 2.1 shows style-transfered using the deep CNN approach. Note that when the original content image (Figure 2.1(a)) is style-transferred (Figure 2.1(c)), the shape of buildings and other structures remains the same whereas color and brush strokes correspond to the style image (Figure 2.1(b)). Figure 2.2 shows artistic style transfer for facial images.

Despite the success of artistic style transfer, facial style transfer remains challenging due to the requirement of photo-realism and semantic consistency. Human vision is very sensitive to facial irregularities and even small distortions can make a face look unrealistic [137, 172]. Figure 2.3(c) demonstrates that when style transfer using CNN is used with a style from a facial image (Figure 2.3(b)), the texture level information like wrinkles do not get transferred in the output image. Similar results are observed when a multiscale style transfer method in [168] is used (Figure 2.3(d)).

In this work, we address the problem of photo-realistic facial style transfer, which transfers *facial texture* from a new style image while preserving most of the original *facial structure* and identity (Figure 2.4). This is joint work with Hang Zhang [90]. Facial texture comprises skin texture details like wrinkles, pigmentation and pores, while facial structure consists of the meso-structures such as eyes, nose, mouth and face shape. Our approach has important implications in commercial applications and dermatology, such as visualizing the effects of age, sun exposure, or skin treatments (e.g. anti-aging, acne).

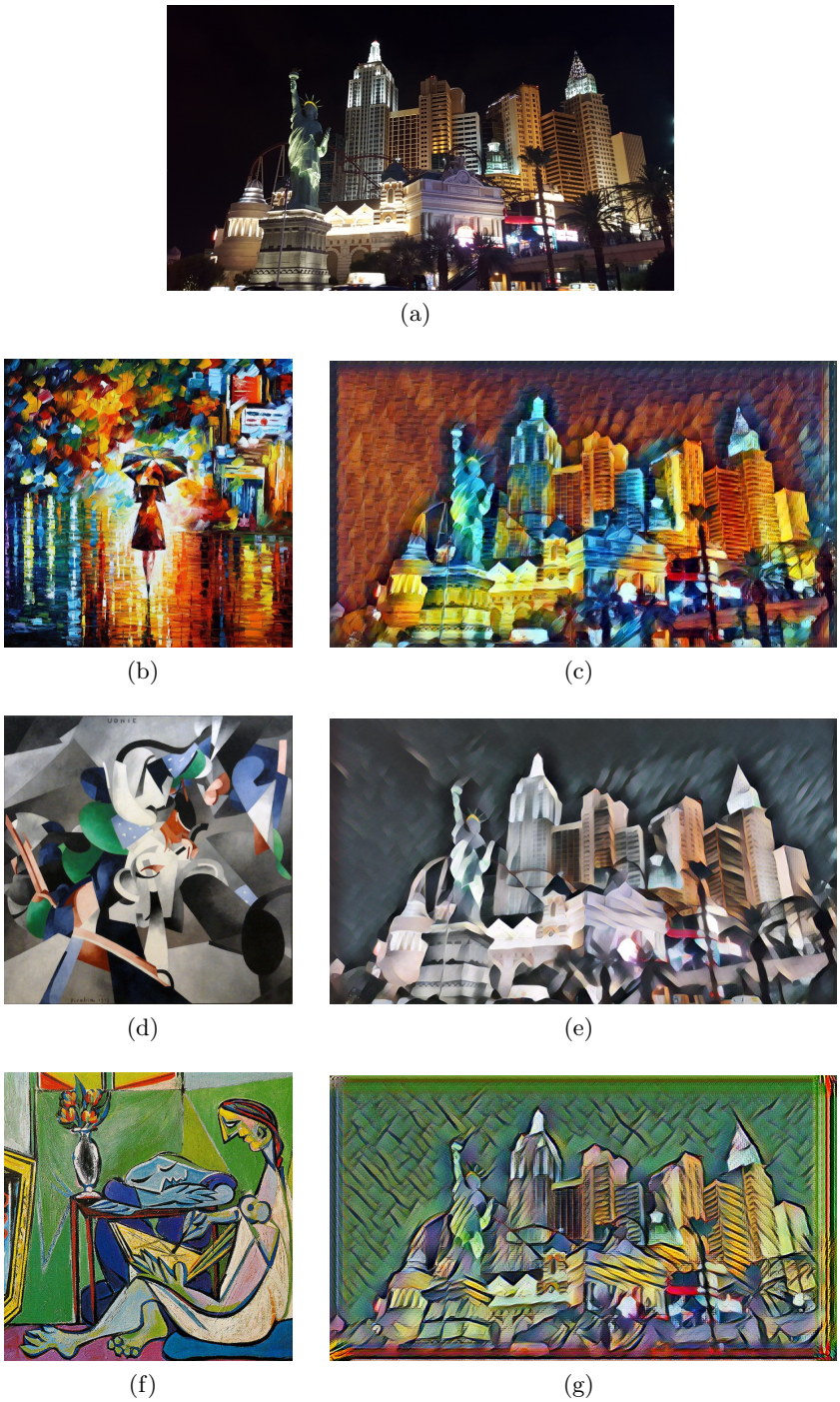
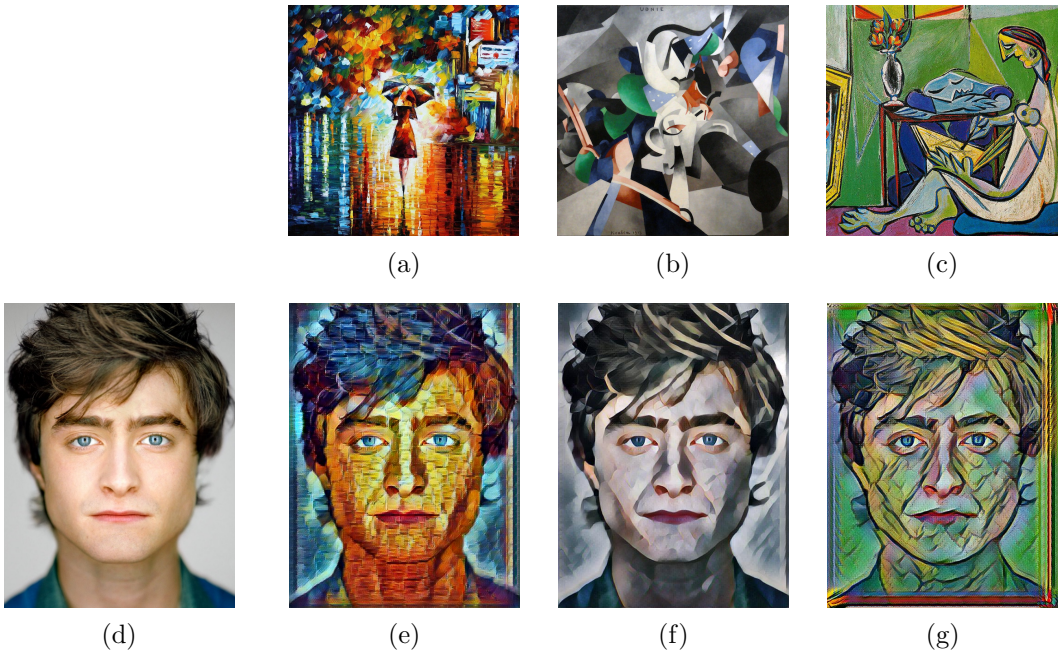


Figure 2.1: **Artistic style transfer for natural images using CNN.** (a) Original Content Image (x_c). [Left] Style Images (x_s): (b) Rain Princess by Leonid Afremov, (d) Udnie by Francis Picabia, (f) The Muse byPablo Picasso. [Right] (c), (e), (g): Style-transfered image (x_t) with contents of image x_c and style of image x_s. Implemented using the fast style transfer code by [44] based on [53, 83, 182]. Content/style photos: Martin Scheoller/Art+Commerce.



(a) (b) (c)

(d) (e) (f) (g)

Figure 2.2: **Artistic style transfer for facial images using CNN.** [Top] Style Images (x_s): (a) Rain Princess by Leonid Afremov, (b) Udnie by Francis Picabia, (c) The Muse byPablo Picasso. [Bottom] (d) Original Content Image (x_c), (e), (f), (g): Style-transfered image (x_t) with contents of image x_c and style of image x_s. Implemented using the fast style transfer code by [44] based on [53, 83, 182]. Content/style photos: Martin Scheoller/Art+Commerce.

Style transfer of artistic work is typically approached by synthesizing a style texture based on the semantic content of the input image [42, 43, 104, 191]. Classic algorithms match the feature statistics of multi-scale representations [31, 76, 150]. Gatys et al. [53, 55] first adopted a pre-trained CNN [171] as a statistical feature representation to provide an explicit representation of image content and style. The output image is generated by solving an optimization problem which minimizes both content and style differences and iteratively passes the gradient directly to the image pixels. The advantage of this method is that an image can be combined with any style image and explicit training is not required. However, it is slow since the optimization is performed at each iteration. Recent work also explores real-time style transfer by training feed-forward networks while approximating the optimization process which outputs the style transferred images directly [83, 115, 182], However, this requires a separate network to be trained for each style. Many variations of the above methods have been proposed to improve the perceptual quality of the style-transfered images. The work in [52] preserves the color of the content image. In [56]

(a) (b) (c) (d)

Figure 2.3: **Style transfer for facial images.** (a) Original Content Image (x_c). (b) Style Image (x_s). (c) Style-transfered image using CNN methods [44, 53, 83, 182]. (d) Style-transfered image using multiscale method in [168]. Note that even though contents from image x_c are retained in output images (c) and (d), the texture level information like wrinkles is not transferred from image x_s. Content/style photos: Martin Scheoller/Art+Commerce.

perceptual attributes like spacial context, scale and color are preserved. Style transfer methods have been extended to multi-style [40, 199] where a single CNN is trained on multiple styles and allows one or more styles to be applied on a content image.

Shih et al. propose a multiscale method to robustly transfer local statistical information from style image to content image [168]. Their goal is to automatically stylize an unprocessed headshot photo. The content and the style image are decomposed using Laplacian pyramids. At each level of the pyramid, energy maps of the Laplacian outputs are computed to account for local variations. A transfer function is computed from the energy maps and used to match the energy distribution of the content and style images. This multi-scale method transfers the local spacial contrast and color distribution from style image to the content image. As shown in Figure 2.3(d), it does not capture the textural information of style image.

Despite the rapid growth of artistic style transfer work, photo-realistic facial style transfer remains challenging due to the need of preserving local semantic consistency while transferring skin texture. The Gram matrix is often used as a gold-standard style representation. Minimizing the difference of a global representation of Gram matrix does not sufficiently enforce local semantic consistency at meso-structures such as lower facial contour, eyes and mouth as shown in Figures 2.6 and 2.7 (last column). A recent method [125] incorporates

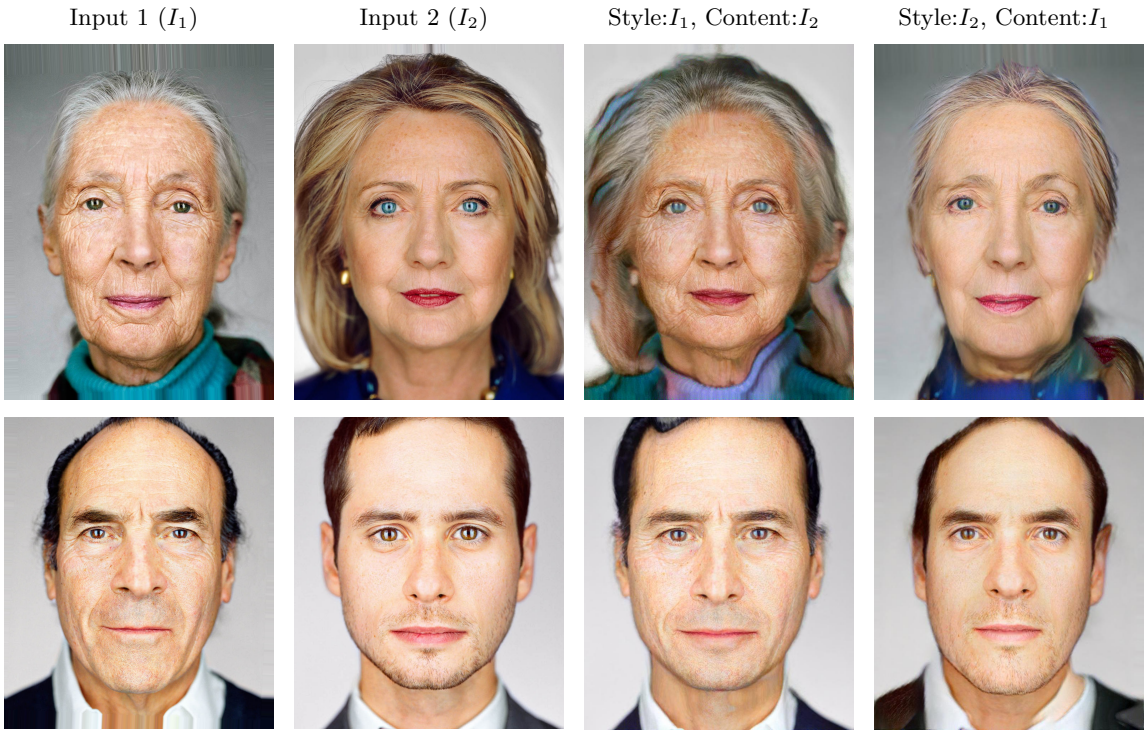


Figure 2.4: **Identity-preserving Facial Texture Transfer (FaceTex).** The textural details are transferred from style image to content image while preserving its identity. FaceTex outperforms existing methods perceptually as well as quantitatively. Column 3 uses input 1 as the style image and input 2 as the content. Column 4 uses input 1 as the content image and input 2 as the style image. Figures 2.6 and 2.7 shows more examples and comparison with existing methods. Input photos: Martin Scheoller/Art+Commerce.

the Gram matrix with semantic segmentation and achieves high quality results for photo-realistic style transfer in scene images. This approach removes distortions in architectural scenes but is not designed for facial texture transfer and has no mechanism for retaining facial structure. Our approach is developed with the specific goal of maintaining the content face identity.

Markov random field (MRF) models have been used widely for representing image texture [201] by modeling the image statistic at a pixel or patch level and the dependence between neighbors. Classic texture synthesis methods using MRF [43] [191] provide new texture instances using an MRF texture model. A recent work called MRF-CNN [115] leverages the local representation of MRF and the descriptive power of CNN for style transfer. However, this method also transfers meso-structures from the style image. For faces, this facial structure sourced from the style image leads to an undesirable change in facial

identity during the texture transfer as in Figures 2.6 and 2.7 (column 3).

As the **first contribution** of this work, we introduce *Facial Semantic Regularization* that consists of a *Facial Prior Regularization* and *Facial Structural Loss* for preserving identity during the texture transfer. Facial identity incorporates facial structure and shapes. We suppress the changes around the meso-structures by introducing the Facial Prior Regularization that smoothly slows down the updating. Additionally, we tackle the challenge of preserving facial shape by minimizing a Facial Structure Loss, which we define as an identity loss from a pre-trained face recognition network that implicitly preserves the facial structure.

The **second contribution** of this work is the development of an algorithm for *Identity Preserving Facial Texture Transfer* which we call *FaceTex* along with a complete benchmark of facial texture transfer with a novel metric for quantitative evaluation. Our approach augments the MRF-CNN framework with the Facial Semantic Regularization and faithfully transfers facial textures and preserves the facial identity. We provide a complete benchmark that evaluates style transfer algorithms on facial texture transfer task. Prior methods typically rely on perceptual evaluation of results, which makes it difficult to quantitatively compare them. We propose metrics that quantify the facial structure consistency as well as texture similarity. The experimental results show that the proposed FaceTex outperforms the existing approaches for identity-preserving texture transfer perceptually as well as quantitatively.

2.2 Methods

2.2.1 Texture Representation

We follow prior work of MRF-CNN [115] for texture representation and briefly describe it for completeness [115]. A pre-trained VGG-19 [171] is used as a descriptive representation of image statistics, and the feature-maps at layer l for input image x is denoted as $\Phi^l(x)$. For a given content image x_c and a style image x_s, the facial texture is transferred from x_s to the output/target image x_t by minimizing the difference of local patches. Let $\Psi(\Phi^l(x))$ denote the set of the local patches on the featuremaps. For each patch $\Psi_i(\Phi^l(x_t))$,

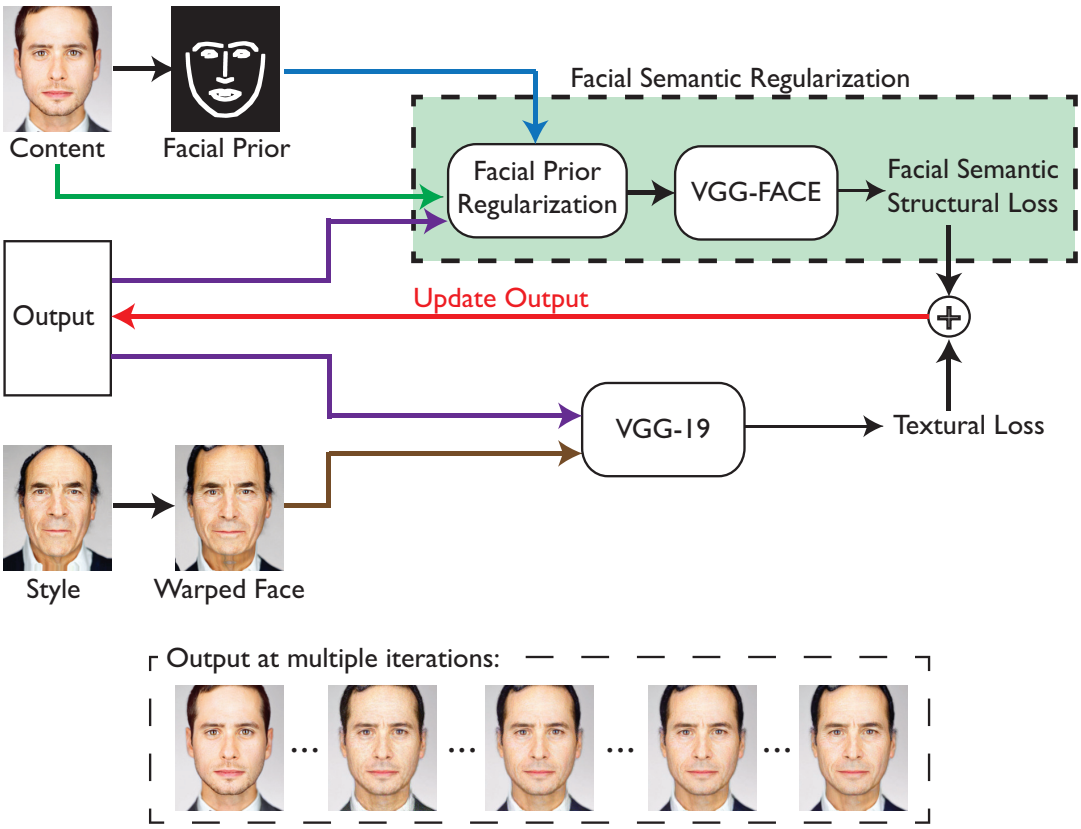


Figure 2.5: **Overview of our method.** Facial identity is preserved using Facial Semantic Regularization which regularizes the update of meso-structures using a facial prior and facial semantic structural loss. Texture loss regularizes the update of local textures from the style image. The output image is initialized with the content image and updated at each iteration by back-propagating the error gradients for the combined losses. Content/style photos: Martin Scheoller/Art+Commerce.

the difference with the most similar patch in the style image $\Psi_{NN(i)}(\Phi^l(x_s))$ (among N_s patches) is minimized. The distance of the nearest neighbor is defined using normalized cross-correlation as

$$NN(i) = \underset{j=1,...N_s}{\operatorname{argmin}} \frac{\Psi_i(\Phi^l(x_t))\Psi_j(\Phi^l(x_s))}{|\Psi_i(\Phi^l(x_t))| \cdot |\Psi_j(\Phi^l(x_s))|}. \tag{2.1}$$

The texture loss is the sum of the difference for all the N_t patches in the generated image and is given by

$$\ell^l_{tex}(x_t, x_s) = \sum_{i=1}^{N_t} \|\Psi_i(\Phi^l(x_t)) - \Psi_{NN(i)}(\Phi^l(x_s))\|^2. \tag{2.2}$$

In contrast to the Gram Matrix that gives global impact to the image, MRF-CNN is good for preserving local textural structures. However, it also carries the semantic information from the style image, which violates the goal of preserving facial identity. For this, we augment the MRF-CNN framework with additional regularizations.

2.2.2 Facial Semantic Regularization

Facial identity consists of meso-structures including eyes, noise, eyebrow, lips and face contour. We tackle the problem of preserving facial identity by suppressing local changes around these meso-structures and minimizing the identity loss from face recognition network, which implicitly preserves the semantic facial structure.

Facial Prior Regularization

Inspired by the dropout regularization [175] which randomly drops some units and blocks the gradient during the optimization, we build a facial prior regularization that smoothly slows down the updating around the meso-structures. For generating the facial prior mask, we follow the prior work [162] to generate 66 landmark points and draw contours for meso-structures. Then we build a landmark mask by applying a Gaussian blur to the facial contour and normalize the output between 0 and 1, which provides a smooth transition between meso-structures and rest of the face. For implementation, we build a CNN layer that performs an identity mapping during the forward pass of the optimization, and scales the

gradient with an element-wise product with the face prior mask during back-propagation.

Facial Semantic Structure Loss

Deep learning is well known for learning hierarchical representations directly from data. Instead of manually tackling preservation of facial structure, we minimize the perceptual difference of a face recognition network to force the output image to be recognized as the same person depicted in the input/content image. VGG-Face [145] is trained on millions of faces and has superior discriminative power for face recognition, which captures the facial meso-structures for identifying the person. Instead of minimizing the final classification error, we minimize the difference of mid-level feature-maps, because the mid-level features are already discriminative for preserving facial identity. Let $\delta^i(x)$ denote the feature-maps at a i-th layer of a pre-trained VGG-Face for input image x. The structure loss is the L^2-distance of the feature-maps and is given by

$$\ell_{face}(x_t, x_c) = \sum_{i=1}^{N_l} \frac{1}{C_i H_i W_i} \|\delta(x_t) - \delta(x_c)\|^2, \tag{2.3}$$

where N_l is total number of layers for calculating structure loss, and C_i, H_i and W_i are the number of channels, height and width of the feature-map, respectively.

2.2.3 Identity Preserving Facial Texture Transfer

Pre-processing

To maintain facial structural consistency and avoid artifacts, we warp the style image to the facial structure of the content image. First, 66 facial landmark points are generated for the content and style images using an existing facial landmark detection algorithm [162]. The style image is then morphed and aligned to the content image [5]. To further align the face contour we apply sift-flow, which uses dense SIFT feature correspondences for alignment while preserving spacial discontinuities [120].

Loss functions

Reconstructing the image from the loss of highly abstracted pre-trained networks makes the image look unrealistic and noisy. We follow the prior work [83, 115, 199] which uses total variation regularization (TV loss) to encourage the smoothness of the output image x, which is given by the squared norm of the gradients:

$$\ell_{TV}(x) = \sum_{i,j} \left((x_{i,j+1} - x_{i,j})^2 + (x_{i+1,j} - x_{i,j})^2 \right). \tag{2.4}$$

We use a weighted combination of texture loss, facial structure loss and TV loss to find the output estimate $\hat{x_t}$ as follows

$$\hat{x_t} = \underset{x_t}{\operatorname{argmin}} \sum_{l=1}^{L} \lambda_{tex}^{l} \ell_{tex}^{l}(x_t, x_s) + \lambda_{face} \ell_{face}(x_t, x_c) + \lambda_{TV} \ell_{TV}(x_t), \tag{2.5}$$

where L is total number of layers for texture loss and λ_{tex}^{l}, λ_{face} and λ_{TV} are the balancing weights for texture loss, facial structure loss and TV loss. The optimization is performed by manipulating the the content image x_c by iteratively updating the image pixels using an L-BFGS solver.

2.3 Experimental Results

2.3.1 Facial Style Transfer Benchmark

Baseline Approaches. We use the publicly available implementation of *Neural Style transfer* for comparison [55, 82]. Gatys et al. [55] generates an output image x_t from content image x_c and style image x_s by jointly minimizing the content loss and the style loss iteratively. The content loss is given by the L^2-distance of the feature-maps at each convolution layers for the output and the content images. The style loss is the Frobenius norm of the Gram matrix difference of the feature-maps of output image and the style images at each layer. The weighted combination of the losses is minimized to obtain the output image

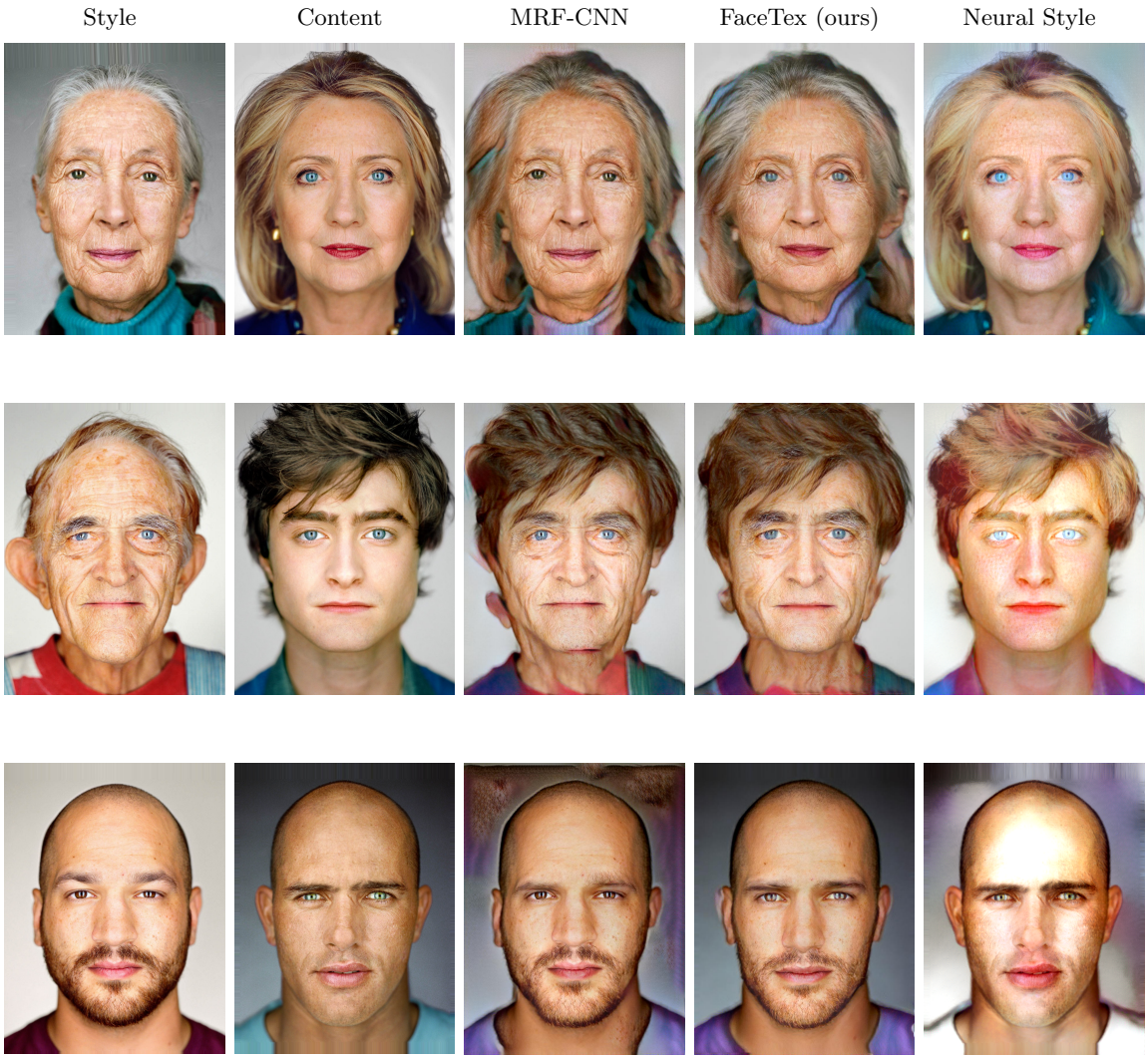


Figure 2.6: **Qualitative Results.** Our facial texture transfer on different content-style pairs. FaceTex (our approach, 4th column) preserves the identity of the content image and also transfers the textural details from the style image. Neural style (last column) preserves the identity but does not transfer the textural details. MRF-CNN (3rd column) transfers the textural details but does not preserve the content image identity as well as FaceTex (compare 3rd and 4th column to content image in 2nd column). Content/style photos: Martin Scheoller/Art+Commerce.

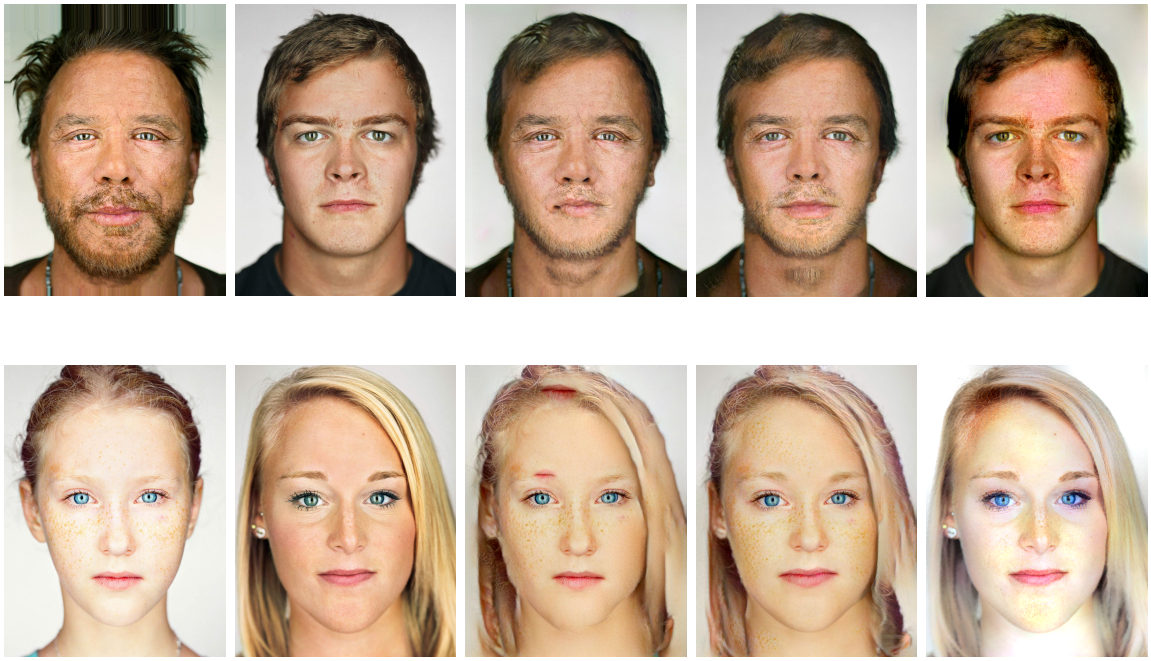


Figure 2.7: **Qualitative Results.** Our facial texture transfer on different content-style pairs. FaceTex (our approach, 4th column) preserves the identity of the content image and also transfers the textural details from the style image. Neural style (last column) preserves the identity but does not transfer the textural details. MRF-CNN (3rd column) transfers the textural details but does not preserve the content image identity as well as FaceTex (compare 3rd and 4th column to content image in 2nd column). Content/style photos: Martin Scheoller/Art+Commerce.

as

$$\hat{x_t} = \underset{x_t}{\operatorname{argmin}} \lambda_c ||\Phi(x_t) - \Phi(x_c)||^2 + \lambda_s \sum_{l=1}^{L} ||\mathcal{G}^l(x_t) - \mathcal{G}^l(x_s)^l||_F^2 + \lambda_{TV} \ell_{TV}(x_t), \tag{2.6}$$

where $\Phi(x_t)$ and the $\Phi(x_c)$ are the feature-maps of output and style images, $\mathcal{G}^l(x_t)$ and $\mathcal{G}^l(x_s)$ are the Gram matrices of the feature-maps of output and style images at layer l; L is the total number of layers; λ_c, λ_s and λ_{TV} are the weights for content loss, style loss and TV loss. In these experiments, we use $\lambda_c = 5$, $\lambda_s = 100$ and $\lambda_{TV} = 10^{-3}$. We use the L-BFGS solver for 1000 iterations. VGG-19 [171] pre-trained network is used for computing feature-maps. Layer relu4_2 is used for content loss while layers relu1_1,relu2_1,relu3_1,relu4_1 and relu5_1 are used for style loss.

We also compare our work with it MRF-CNN [115]. The output image is generated by minimizing the patch difference with the style image and preserving the high-level structure the same as in the content image. The loss function consists texture loss, content loss and TV losses:

$$\hat{x_t} = \underset{x_t}{\operatorname{argmin}} \lambda_c ||\Phi(x_t) - \Phi(x_c)||^2 + \lambda_s^l \sum_{l=1}^{L} \ell_{tex}^l(x_t, x_s) + \lambda_{TV} \ell_{TV}(x_t), \tag{2.7}$$

where $\ell_{tex}(x_t, x_s)$ is the texture loss as in Section 2.2.1, λ_c, λ_s and λ_{TV} are the wights for content loss, style loss and TV loss. Layers relu3_1 and relu4_1 of VGG-19 are used for texture loss and layer relu4_2 for content loss. Neural patches of size 3×3 are used to find the best matching patch. Three resolutions with 100 iterations each are used.

Implementation Details. We follow the work of MRF-CNN using layers relu3_1 and relu4_1 of VGG-19 [171] for texture loss. Layer relu4_2 of a pre-trained VGG-Face [145] is used for facial semantic structure loss. The facial prior mask is generated by connecting the landmark points using 40 pixel thickness line and applying a Gaussian blurring with the kernel size of 65 and standard divination of 30. In addition, the background mask provided in the dataset is also used. We incorporate facial prior regularization to block the changes of facial prior and background regions. We resize the content and style images to $1,000$ pixels along the long edge. The output image is initialized with the content image

and the optimization is performed using the L-BFGS solver. We follow Li and Wand [115] using a multi-resolution process during the generation, the content and style images are scaled accordingly. We start with $\frac{1}{4}$ resolution and scale up by a factor of 2, and perform 200 iterations at each resolution. We use the same resolution for both baselines and our approach in this experiment.

Metrics for Quantitative Evaluation. We identify two metrics to quantitatively measure the facial structural inconsistency and texture similarity of the output image x_t with the content image x_c and the style image x_s .

Landmark Error: Using the methods described in section 2.2.3, we obtain $L = 66$ landmarks for each facial image. The output image has same facial structure if its landmark points remain the same as content image. The mean square error of the landmarks between the two images accounts for the facial structural inconsistency between them. Lower error indicates identity is preserved. The landmark error between two facial images is given by the L^2 distance of the pixel coordinates for the landmark points.

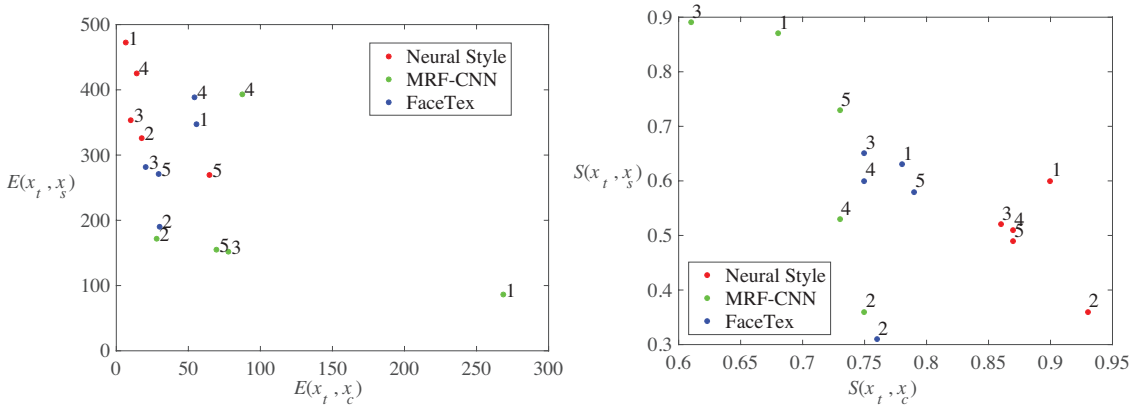
Texture Correlation: To measure the similarity between the output image and the input images, we can extract skin patches from the images and use the *normalized correlation coefficient*. Higher value of correlation coefficient indicates a better match of facial textures. Texture similarity of two patches p and q is given by:

$$S(p,q) = \frac{\sum_{ij}(p_{ij} - \bar{p})(q_{ij} - \bar{q})}{\sqrt{\sum_{i,j}(p_{ij} - \bar{p})^2 \sum_{i,j}(q_{i,j} - \bar{q})^2}}, \tag{2.8}$$

where p_{ij} and q_{ij} are the image values at pixel coordinates (i, j) of the patches, $\bar{p}$ and $\bar{q}$ are the average pixel values of patches p and q.

2.3.2 Qualitative and Quantitative Comparison

We use the head portrait dataset provided by Shih *et al.* [168] for evaluation. Figures 2.6 and 2.7 shows the comparison of the output image generated using FaceTex with Neural Style and MRF-CNN. We provide additional comparison in the supplementary material. We observe that Neural Style preserves the facial structure and shape well but fails to transfer the texture, which demonstrates that the Gram Matrix transfers global styles well



(a) (b)

Figure 2.8: **Quantitative evaluation** for each content-style pair. (a) Landmark Error. For FaceTex, $E(x_t, x_c)$ is small and much closer to Neural Style than MRF indicating that it preserves identity as in Neural Style approach. (b) Texture similarity. For FaceTex, $S(x_t, x_s)$ is high and closer to MRF-CNN, transferring tecture from style image.

but fails to preserve the local finer texture and also makes the image unrealistic. MRF-CNN transfers local texture very well but it does not preserve the meso-structures which leads to more significant change the observed facial identity. Our proposed FaceTex approach generates photo-realistic images and outperforms all the baseline approaches in transferring facial texture as well as preserving the facial identity.

	Neural Style	**MRF-CNN**	**FaceTex (ours)**
$E(x_t, x_c)$	22.61	106.17	37.93
$E(x_t, x_s)$	369.39	191.47	295.93
$S(x_t, x_c)$	0.89	0.70	0.76
$S(x_t, x_s)$	0.47	0.68	0.55

Table 2.1: **Metrics for quantitative evaluation.** The average metric values of the pairs in Figures 2.6 and 2.7 are reported here. For FaceTex, landmark error between output and content $E(x_t, x_c)$ is much lower than MRF-CNN indicating it is better at preserving identity. Texture similarity between output and style $S(x_t, x_s)$ is higher in FaceTex than Neural Style which shows that it is better in transferring texture.

The quantitative comparison matches the conclusion of qualitative observation, and the results of landmark and texture metrics are listed in Table 2.1. The average values of different metrics are reported for the content-style pairs in Figures 2.6 and 2.7. $E(x_t, x_c)$ and $E(x_t, x_s)$ are the landmark errors of output image with content and style images, respectively. $E(x_t, x_c)$ is very low for Neural Style (22.61) but high for MRF-CNN (106.17)

indicating that identity is preserved by the Neural Style approach. For FaceTex, $E(x_t, x_c)$ is much lower (37.93) than MRF-CNN and preserves the identity. In contrast, higher value of $E(x_t, x_s)$ indicates that facial structural similarity is not maintained with the style image as expected. $S(x_t, x_c)$ and $S(x_t, x_s)$ are the texture similarities of output image with content and style images, respectively. For each content-style pair, we extract three patches and report their average normalized correlation coefficient as the texture similarity of the image. These patches are 100×100 and in forehead and both cheek regions, localized by face structure landmarks. For texture transfer, a large value of $S(x_t, x_s)$ indicates that texture is successfully transferred to output image from the style image. $S(x_t, x_s)$ is highest for MRF-CNN (0.68) and lowest for Neural Style (0.47), whereas for FaceTex similarity lies between MRF-CNN and Neural Style (0.55).

Figure 2.8(a) and (b) shows the landmark errors and texture similarity for each of the five content-style pairs in Figures 2.6 and 2.7. Both the error and the similarity measures for FaceTex (blue dots) lie between Neural Style (red dots) and MRF-CNN (green dots), and generally much closer to MRF-CNN.

Ablation Experiments. Figure 2.9 exemplifies the necessity of augmenting the existing methods with multiple regularizations. If only the facial prior regularization is used, the generated output face still looses identity and has artifacts. Adding the facial semantic structure further preserves the identity and suppresses some artifacts.

Limitations. Our method achieves superior performance in identity preserving facial texture transfer and generates photo-realistic images, but still has its limitations. First, our approach is an optimization-based approach, which takes several minutes generating a new image, which limits the applications in real-time. This could be potentially addressed in the future work by combining a feed-forward network and a face alignment network that run in real-time. Second, the texture modeling using MRF-CNN requires high semantic similarity between two input images, which may lead to some unappealing artifacts for mismatches.

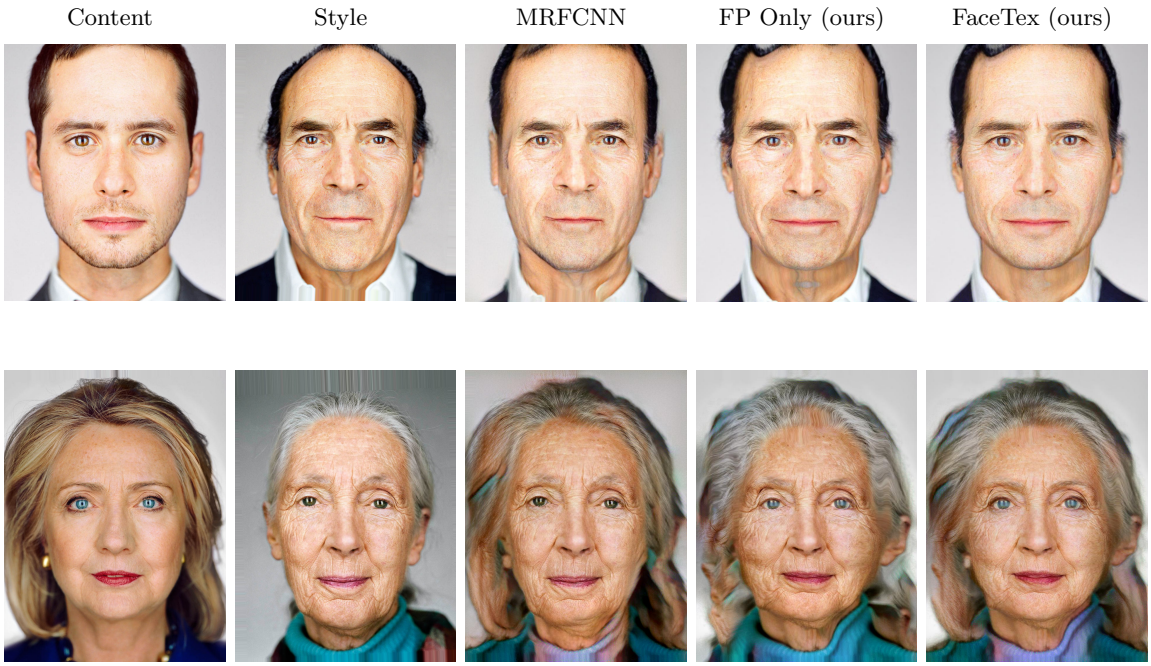


Figure 2.9: **Abalation Experiments.** The effects of facial prior (FP) regularization and facial semantic structure loss. Using FP regularization (column 4) preserves better meso-structure of the faces comparing to MRF-CNN (column 3). Facial semantic loss effectively preserve the facial structure for identity preserving as shown in the last column. Content/style photos: Martin Scheoller/Art+Commerce.

2.4 Conclusion

We have presented the method FaceTex for photo-realistic facial style transfer. By augmenting prior work of MRF-CNN with a novel regularization consisting of a facial prior regularization and the facial semantic structure loss, we can transfer texture realistically while retaining semantic structure so that the identity of the individual remains recognizable. Our results show substantial improvement over the state-of-the-art both in the quality of the texture transfer and the preservation of the original face structure. Quantitative metrics of texture transfer and face structure are also improved using this approach.

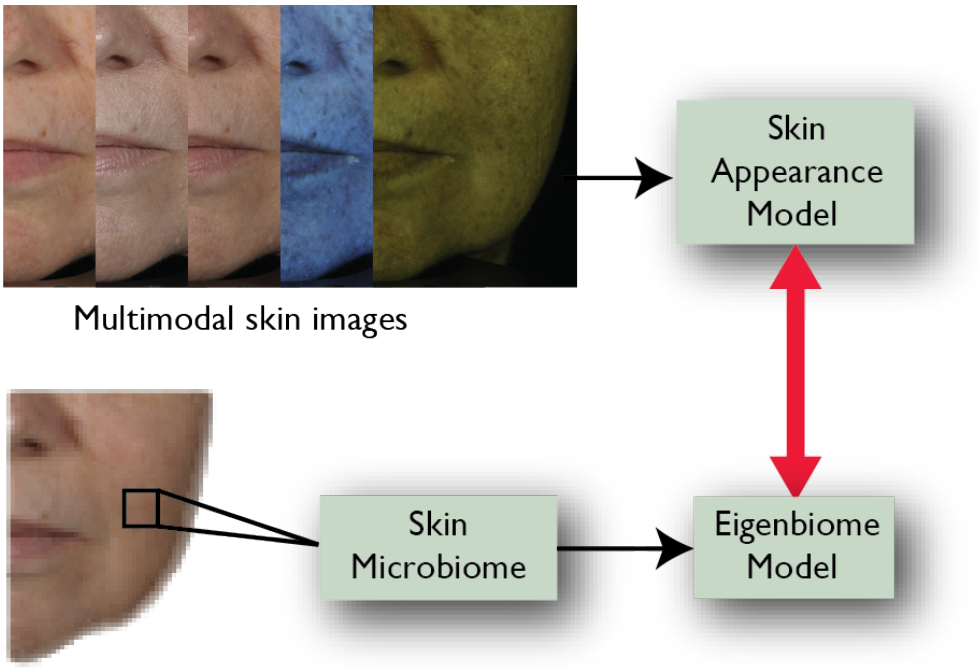
Chapter 3

Skin Appearance and Skin Microbiome

Recent advances in measuring and analyzing the skin microbiome through gene sequencing is revolutionizing our understanding of skin appearance. With skin microbiome measurements, causative relationships between microbes and macro-appearance can be explored. However, gene sequencing is very cost-prohibitive and time-consuming. An exciting opportunity exists to use *photographic imaging and appearance modeling to infer the skin microbiome*, i.e. to effectively "see" the microbiome of a human subject by analyzing skin surface patterns. The pioneering work of [48] shows that healthy skin may harbor a particular strain of benevolent bacteria. The appearance of human skin in this study was divided manually into only two simple classes of good and bad skin appearance. By developing computational models of skin appearance, we provide a more fine-grained quantitative categorization of human subjects using multiple classes of appearance. While prior work in microbiomics shows the association of bacteria with skin appearance [48,69,96], there is no mechanism for automatic inference of the skin microbiome from images. This association of visual patterns to bio-patterns on the skin surface is a novel area that has not been explored. Moreover, while the appearance groups can be verified visually, the groups in high-dimensional microbiome space are latent groups, which are difficult to be evaluated by a human oracle. We call this problem of discovering this latent grouping associated with appearance groups as *causative appearance analysis*.

In this chapter, an approach that uses multi-modal skin imaging and sparse coding to link microbiome to skin texture is presented (Figure 3.1). Then the framework of *appearance-driven multiview co-clustering - AMCO* is introduced, which is the ongoing work and has a powerful component of discovery for causative appearance analysis (Figure 3.3).

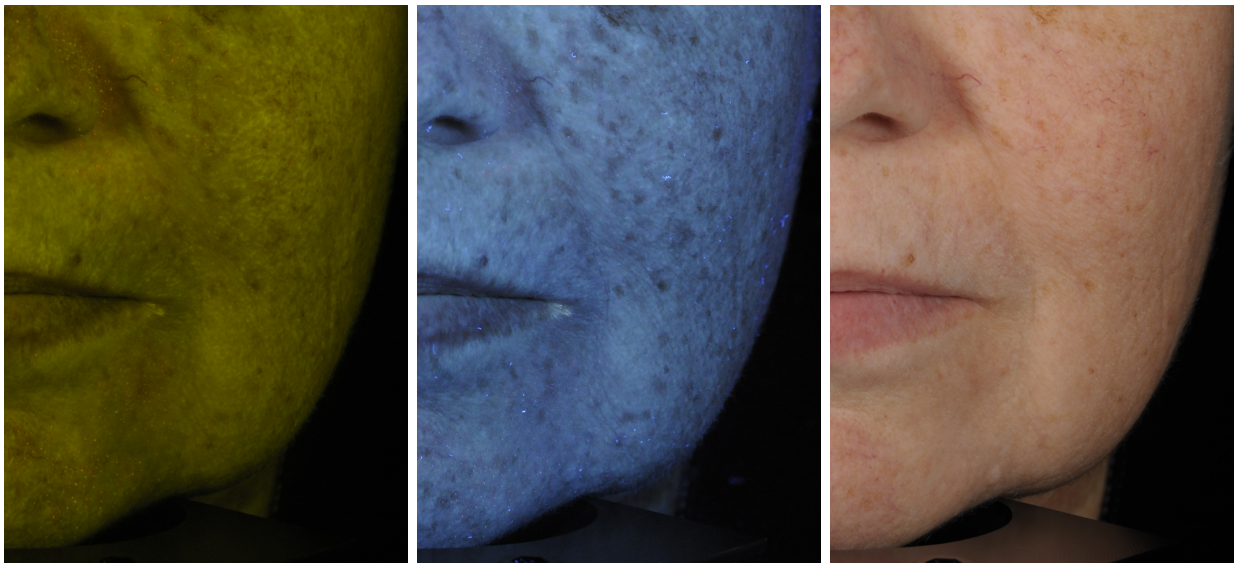
For computational skin texture, we use a texton-based approach with a neural network



(a)

Figure 3.1: **Appearance Modeling.** The computational skin appearance model characterizes multimodal images based on their attributes using a texton-based approach. The eigenbiome model projects the skin microbiome to a lower dimensional subspace. By identifying overlapping groups in appearance and microbiome clusters, we show that our skin appearance model is predictive of the underlying microbiome clusters.

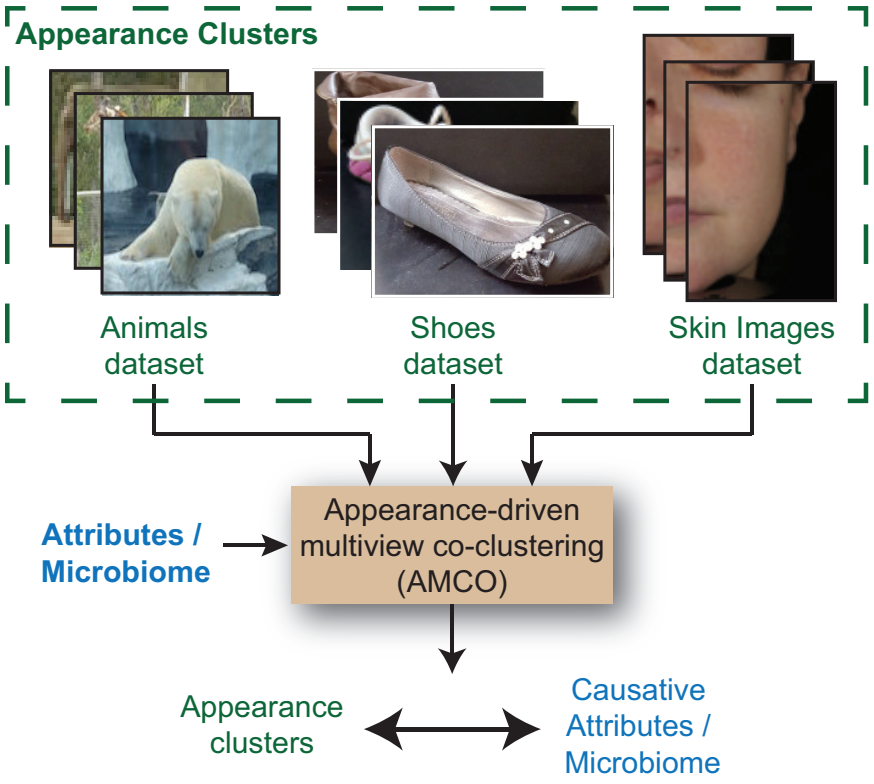
classifier to categorize skin regions based on the distribution of known attributes. The eigenbiome model projects the skin microbiome to a lower dimensional subspace. Projections of microbiome using non-negative matrix factorization (NMF) reveal a physically realizable eigenbiome where the eigenvectors are all positive components and represent particular concentrations of microbes. For our experiments, we capture appearance measurements from 48 human subjects with multimodal images: fluorescence excitation with blue-light (FLUO), fluorescence excitation with ultraviolet radiation (UV), cross polarization (XPOL), parallel polarization (PPOL) and visible light (VISI) (Figures 3.2 and 3.4). Sequencing of swabs from the forehead skin of the 48 subjects gives the corresponding skin microbiome. Using both eigenbiome and skin texton modeling, we have identified overlapping groups in appearance and microbiome and shown that our skin appearance models are predictive



(a) (b) (c)

(d) (e)

Figure 3.2: **Facial images of a subject captured in different modalities.** (a) Fluorescence excitation with blue-light (appears green). (b) Fluorescence excitation with UV radiation (appears blue). (c) Cross-polarization. (d) Parallel-polarization. (e) Visible light. Fluorescence excitation captures skin textures not revealed by the visible light. The most salient attributes are apparent in FLUO and UV images (see Figure 3.6), therefore these modalities are used for computational skin modeling.



(a)

Figure 3.3: **AMCO Framework.** Appearance-driven multiview co-clustering (AMCO) that finds associations between appearance space and a secondary space.

of the underlying microbiome clusters. Therefore, convenient and instantaneous multi-modal photography may be sufficient for inferring microbial characteristics. This proposed methodology is the first of its kind to use a computational model of skin macro-appearance to predict the microbiome clusters.

After establishing the link between skin appearance and microbiome, we can pose and potentially answer the question, *Why do skin surfaces look the way they do?* By discovering the latent grouping associated in the secondary microbiome space with appearance groups, we can discover potential causative associations.

Two algorithm groups in prior work are useful for discovering this latent grouping. However, both groups have limitations. First, algorithms for *multiview clustering* enable interdependent clustering in two different feature spaces [7, 178, 196]. The term multiview refers to multiples views of the subject typically in two feature spaces obtained from heterogenous sources (e.g. appearance features and manufacturing parameters). Clusters in

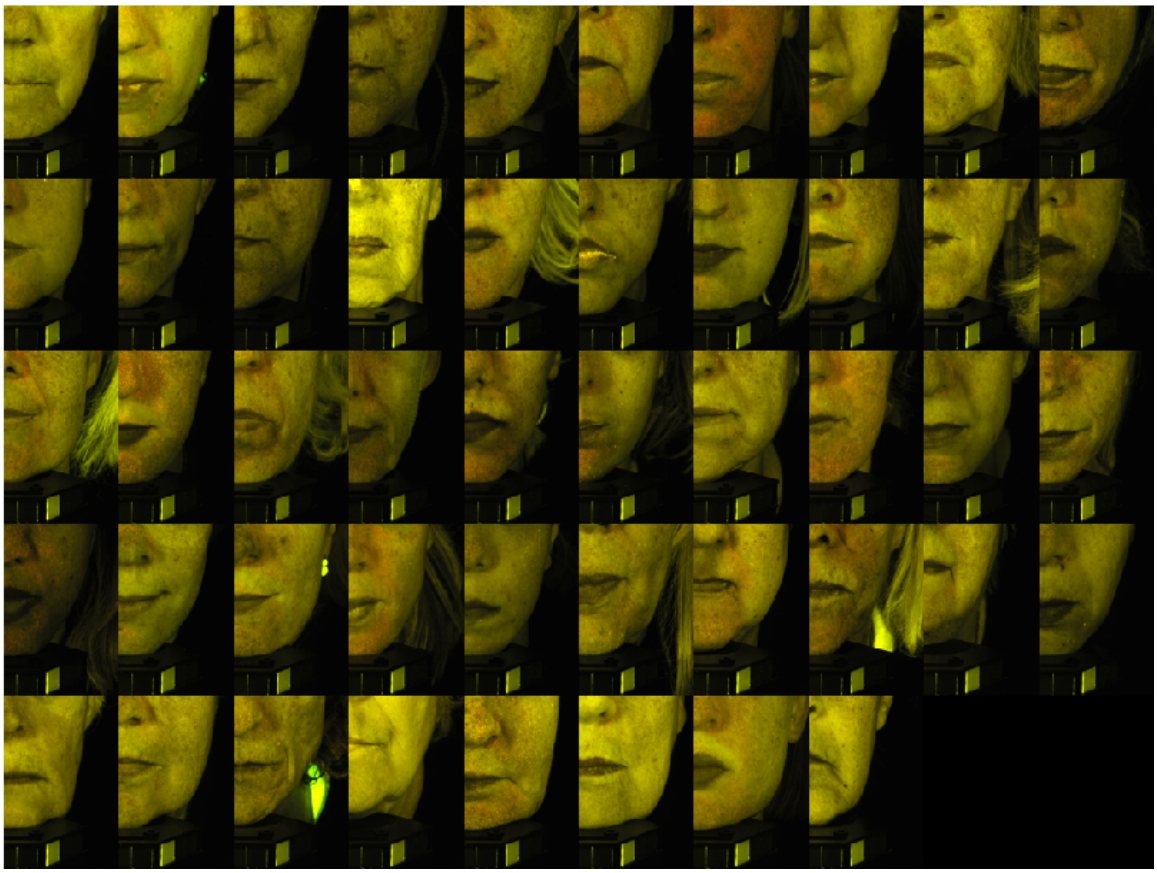


Figure 3.4: **Appearance-microbiome Dataset.** Partial faces of subjects imaged in fluorescence excitation with blue-light (FLUO) modality. Our database contains 48 subjects with age varying between 25 and 68. The faces of each subject are imaged from left, right and frontal views under five modalities. Each image is 4032 × 6048 pixels in size. The forehead regions used in the experiments are from the frontal views. Sequencing of swabs from the forehead skin of the 48 subjects gives the corresponding skin microbiome.

two spaces are formed simultaneously so that there is an agreement between the grouping in each space (Figure 3.5(a)). Without multiview clustering, a naive way to deal with two feature spaces is to cluster independently in each space; but this approach can result in cluster inconsistency where subjects/objects may be assigned to different clusters in each space. Consider how this is problematic; if a group of objects have the same appearance group, these objects should also belong to same process space group. In this manner, the cluster agreement is an association of the appearance with process parameters. This consistency of group members is fundamental in multiview clustering methods and often appears as an agreement term in the objective function [70, 166, 188]. Another naive approach is feature concatenation from different views followed by traditional clustering algorithms like

k-means. However, since the multiple views are obtained from heterogeneous sources, concatenating can result in scaling issues and over-fitting [196]. Often concatenated clustering results are poor as compared to the clustering results using a single-view feature set [9]. To effectively combine the information from multiple views, the multiview clustering enforces agreement between clusters so that similar subjects are assigned to the same cluster in each space.

The second group of algorithms we leverage is co-clustering. Co-clustering or biclustering is different from ordinary clustering because it reveals not only the subject groups but also the feature set which causes the grouping (Figure 3.5(b)). It allows simultaneous clustering of data points (rows) and features (columns), identifying groups among the subjects and groups among the features that are associated with that particular subject group. For instance, if animals are clustered using attributes, one might expect the animals in a cluster to be *elephant, rhino, hippo*; but co-clustering would also give a grouping on the features (in this case attributes) that might consist of *grey, large, mammal*.

Our new approach for *appearance-driven multiview co-clustering (AMCO)* groups in two spaces uses transfer-learning dimensionality reduction that creates a new feature space amenable to multiview spectral co-clustering (Figure 3.5(c)). This framework is a pioneering step to discover how the microbiome affects skin appearance by investigating causative links to microbiome. It can also be utilized in other domains to use appearance-based recognition to discover or infer new groups in associated non-visual domains. With recent advances in computer vision, object and texture recognition is quickly becoming a success story. There are many processes that *affect* appearance such as weathering, manufacturing, agricultural, and biological processes. For example, in a manufacturing process the BRDF (appearance) of paints and pigments can be a function of measurable parameters like humidity, machine setting, temperature, curing time and operator characteristics. In fiber science, texture of fabrics (appearance) can be associated with measurable parameters like thickness, friction, wrinkling, and conductivity. AMCO addresses the problem of causative appearance analysis to create a bridge from computer vision appearance-based recognition algorithms to other disciplines that affect and cause fine-grained appearance categories. We demonstrate the utility of AMCO on skin appearance with microbiome dataset in FLUO, UV and XPOL

imaging modalities. In addition, we include the results from an image-attribute datasets: animals with attributes.

The rest of this chapter is organized as follows: section 3.1 provides an overview of the related work in skin modeling and multiview clustering. In section 3.2, the methods for multimodal skin imaging, computational appearance modeling, eigenbome modeling and the AMCO framework are presented. Following this, section 3.3 discusses the results on appearance-microbiome dataset. Finally, section 3.4 concludes and discusses the future work of this research.

3.1 Related Work

Human skin is a complex, multi-layered structure, which hosts various microbial communities. Studies of the skin microbiome [16,68,142] show dependence on genetics, environment and lifestyle as well as a variation over time. The skin microbiome varies according to the location on the body and from individual to individual. While the benefits of gut microbes are well known, knowledge of the skin microbiome is at an early stage [48,69,96].

Prior applications of skin modeling include studies of skin aging [19,105,138], computer-assisted quantitative dermatology [128,170], and lesion classification [97,130]. Interaction of skin's surface and subsurface with light leads to varied skin appearances. Several imaging techniques have been developed in dermatology for analyzing skin health. These imaging techniques include polarized imaging to enhance surface and subsurface skin features, and fluorescence imaging to capture features which are not visible [21, 25, 26, 86, 95, 127]. The images captured using different methods are collectively called *multimodal images*. Multimodal high-resolution skin imaging captures fine scale features like pores, wrinkles, pigmentation. Computational methods have been developed to automatically detect skin conditions like inflammatory acne, erythema and facial sebum distribution from multimodal images [21,73,86]. However, the skin appearance has not been characterized as a collection of quantifiable features using these imaging techniques.

Research in psychophysics established that some basic black and white basic image features are discriminable from other features by pre-attentive human perception [85]. These

basic image features are the elements of textual perception and referred to as *textons*. In computer vision, textons have been used for texture classification and object recognition. A classic approach is to use filter responses or the joint distribution of intensity values over pixel neighborhoods to identify the basic texture elements or textons [22,28,30,113,121,183]. An image can then be represented by a distribution of textons. In [118], a multi-layered approach based on multilevel PCA and multiscale texon features is applied for face recognition. The bag-of-features representation of images using local interest features has been used for object or scene recognition [47, 194], image segmentation [89] and image retrieval [200]. Skin exhibits 3D texture and the appearance varies significantly depending on illumination or viewing direction. Methods have been developed to model skin appearance to account for this variation [24, 27, 113, 132]. Skin reflectance models have also been developed to acquire and render human skin [38, 58, 99, 193]. Local appearance has been linked to attributes [2,146,159], pose [131,148,197] and motion [141,187,195]. In this work we develop an appearance model that can be demonstrably linked to microbiome measurements.

Multiview clustering algorithms combine features from heterogeneous sources to cluster similar subjects in groups and have applications in image retrieval, bioinformatics and text mining. We concentrate on the type of multiview datsets that provide distinct object information from an unrelated source operating in a different space of measurements. We define this problem as distinct from multimodal or multisensor measurements obtained on the same or similar image coordinates. For example, RGB/depth, infrared/visible, MRI/CT are all multimodal examples where complementary information is measured on the same spatial grid coordinates. In some prior work, the term multiview refers to different feature sets extracted from the same image. For example, multiview clustering algorithms [8, 101, 102, 188] are demonstrated on UCI handwritten dataset, where the Fourier coefficients and profile correlations are considered as two different views of the data. Similarly, in [9], SIFT, HOG and GIST feature vectors are extracted from the same image and considered as different views of the dataset. In some prior work, multiview also means different geometric viewing or illumination angles, but this is a different context for the term multiview. For our work, the secondary view has information on the same subjects (or objects) but from a distinct heterogeneous source. There is no common coordinate frame and no alignment

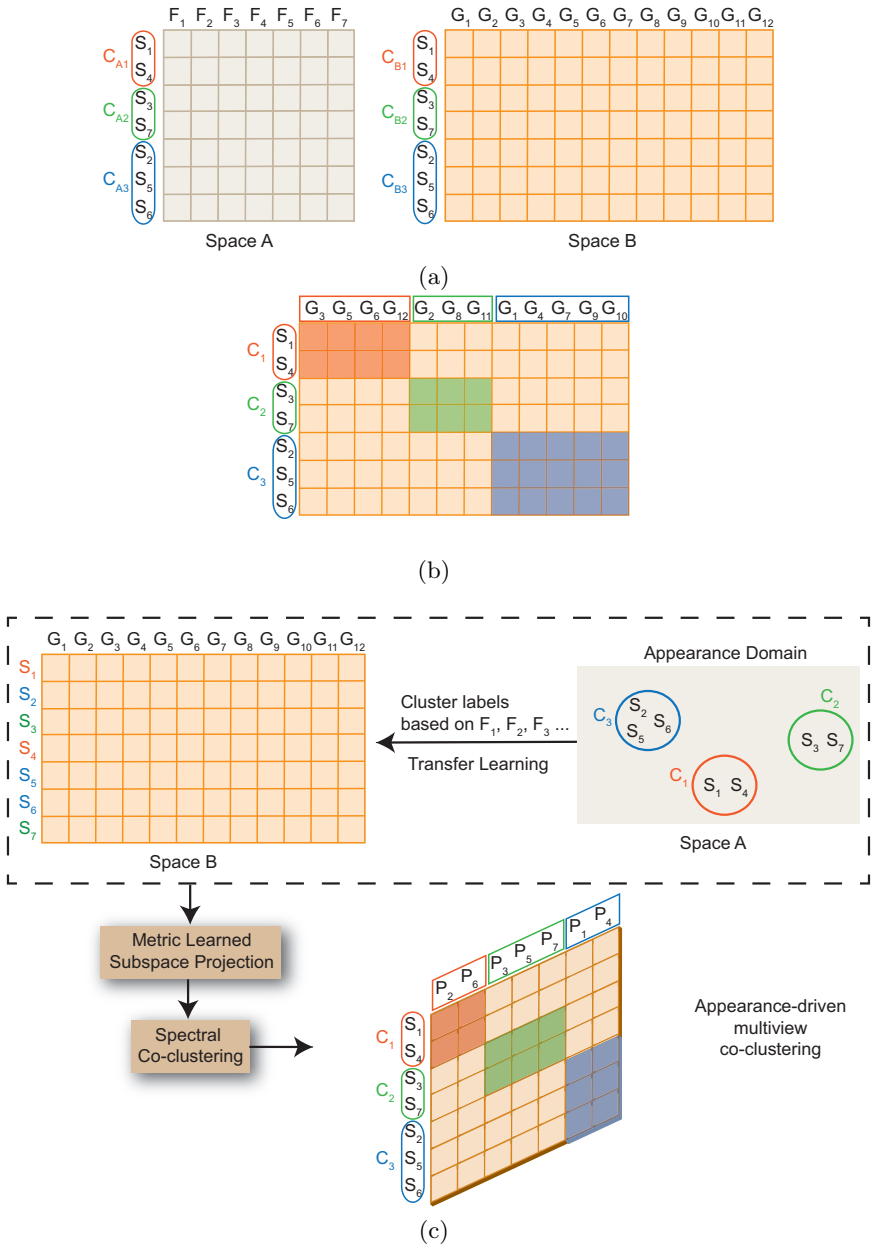


Figure 3.5: **Comparison of multiview clustering, co-clustering, and AMCO.** (a) Multiview clustering: Clusters in two domains are formed simultaneously so that there is an agreement between the grouping in each space. (b) Co-clustering: data points (rows) and features (columns) are clustered simultaneously, grouping together a subset of data points and features with similar behavior. For example, group C_1 with subjects S_1 and S_4 are associated with features G_3, G_5, G_6, G_{12}. (c) Appearance-driven multiview co-clustering (AMCO): Proposed methodology to link groups in appearance space A to features in the secondary space B. For example, groups C_i are obtained by clustering $(F_1, F_2, \ldots)$ in appearance space A. The cluster labels are transferred to space B and supervise MCML for subspace projection of features $(G_1, G_2, \ldots)$. Spectral co-clustering in this new space reveals groups of features associated with clusters C_i such as (P_2, P_6) to C_1.

	multiview	spectral	co-clustering
MV-joint NMF [122], MV-CCA [14, 63, 166, 177], DAKM [84]	X		
MV-kmeans [8], HSF [119]			
multipartite graph [33], MV-cot [102], MV-coreg [101], SDAKM [84], SSVD [112], DiMSC [12]	X	X	
MV-SSVD [176], AMCO	X	X	X

Table 3.1: **Multiview clustering algorithms.** Multiview co-clustering has received only sparse attention in the literature.

can be made since no one-to-one correspondence exist. Examples of such sets include images/words, images/bio-data and images/attributes.

Spectral methods have been widely used for multiview clustering [32, 33] with bipartite graphs constructed from nodes of both views, edges connect nodes from one view to another. Affinity matrices of each view are combined to form an affinity matrix for the multiview bipartite graph to minimize the disagreement between both the views. In multiview co-training (MV-cot) [101], the spectral embedding from one view is used to modify the graph structure of the similarity graph in the other view by projecting the similarity vectors along the first k-eigenvectors of the Laplacian of the similarity matrix. Co-regularized multiview spectral clustering (MV-coreg) [102] is based on the hypotheses that eigenvectors have high pairwise similarity across two views. Pairwise co-regularization minimizes the disagreement between the clustering of two views by minimizing the cost-function obtained using similarity matrices of both views. The joint optimization problem is framed to maximize the matrix trace in both the views as well as the disagreement between the views. A common theme in many multiview methods is an agreement term that is maximized as part of the objective function. Often this agreement term is expressed using regularization or projection to a common subspace. In centroid-based co-regularization, the eigenvectors for each view are regularized towards a consensus (common centroid) eigenvector. The multiview clustering criterion in [122] minimizes the reconstruction error for each view while simultaneously regularizing each view towards a consensus.

Canonical correlation analysis (CCA) for multiview clustering [14, 63, 177] projects descriptor variables in both spaces to a lower dimensional subspace in order to maximize correlation. In [84], a Dual Assignment k-Means (DAKM) algorithm is introduced to cluster videos using human actions and contextual scene information and the mutual information between two clustering results is maximized. The traditional CCA methods have been extended by [63] to incorporate a third view comprised of image semantics of key words. Generalized multiview analysis [166] extends CCA to a supervised framework providing a joint optimization method for the two views. Related work in cross-modal matching [190] learns two lower-dimensional projection matrices to map data in both spaces into a common feature space. Feature selection is done by imposing l_{21}-norm penalties so that discriminative features from the two spaces are coupled. Our method has similarities to this approach except that the projection matrix is learned only for the secondary space in order to match the appearance clusters.

Our approach is different from prior multiview work in that a common subspace for the two views is not used. Instead a subspace in the secondary space is learned that is tuned to the clusters in the primary space. The transformation is learned using maximally collapsing metric learning (MCML) which provides a method to create a new subspace where standard clustering results in clusters consistent with the primary space. That is, all the subjects in the secondary space cluster have the same appearance description.

While prior multiview clustering methods have been shown to improve clustering of data points in multiple views, most do not emphasize grouping of features in each of the views, i.e. feature selection. Unlike multiview clustering, co-clustering addresses this problem of feature selection. For example, in document clustering, co-clustering groups similar documents *and also* groups words which are linked to a particular document group [35]. Similarly, in movie recommender systems, co-clustering of subjects' movie preferences reveals which users have similar movie preferences and the list of those movie preferences [57]. In bioinformatics, data points are clustered according to their gene expression profiles and the subset of genes for each category [129] is revealed in co-clustering. Note that these examples are all single-view co-clustering examples.

We address both problems of spectral co-clustering and multiview analysis. As shown

in Table 3.1, there are many existing methods in the literature for multiview and spectral multiview clustering. However, multiview co-clustering has received only sparse attention in the literature. Co-clustering in multiple views is recently is addressed by [176], which extends the algorithm of [112] based on sparse singular value decomposition (SSVD). However, this work is limited to only two classes. Our method provides multivew co-clustering for multiple classes in an appearance-driven approach.

3.2 Methods

3.2.1 Multimodal Skin Imaging

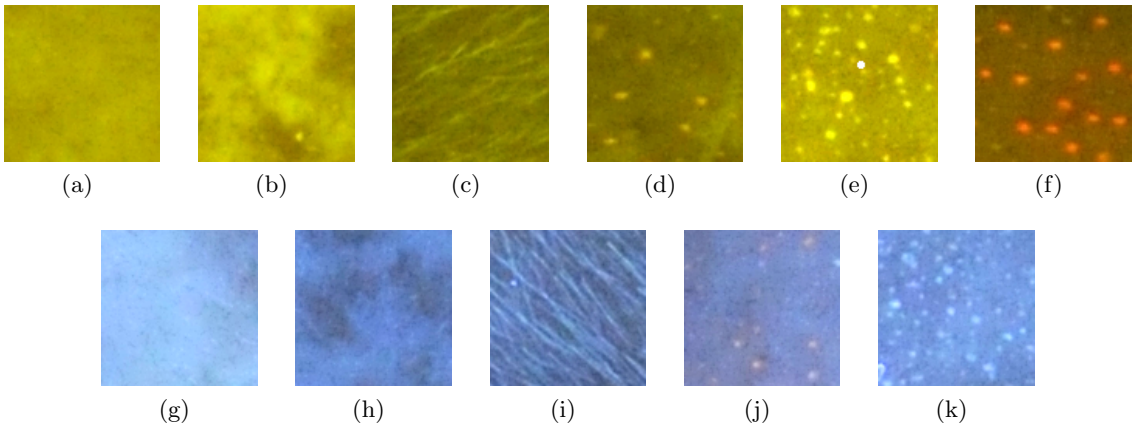


Figure 3.6: **Example patches on facial skin.** The corresponding attribute labels are as follows: In FLUO images: (a) smooth, (b) blotchy, (c) fine hair, (d) sparse sebum dots, (e) dense yellow sebum dots, (f) dense red sebum dots. In UV images: (g) smooth, (h) blotchy, (i) fine hair, (j) sparse sebum dots, (k) dense sebum dots.

Light reflected from the skin constitutes reflection from the stratum corneum (skin's uppermost layer) and light backscattered by the tissues in the dermis which captures skin's surface and subsurface features, respectively [151]. Moreover, not all the skin features can be seen in visible light [95]. Using these characteristics of skin, specialized techniques have been developed to aid dermatologists to capture the skin attributes in different modalities. Polarized light and fluorescence imaging can be used to obtain specific skin features and collectively provide an overall assessment of the skin appearance [95].

Fluorescence excitation with blue light (FLUO) or ultraviolet-A radiation (UV) is used to excite skin elements like keratin, collagen cross-links and elastin cross-links [73,95]. These

skin structures result in image features, but not in visible light. Noticeable skin features include red or yellow dots in FLUO images (Figures 3.6(e) and 3.6(f)) and blotches in UV images (Figure 3.6(h)). Red dots in FLUO images are due to excitation of porphyrins in the pores. Porphyrins are known to be produced by bacteria such as Propionibacterium acnes residing in sebaceous glands. Yellow dots in FLUO images are produced by excitation of "horn" in pores, which is a mixture of keratinocyte ghosts from the sebaceous glands lining, sebaceous lipids, sebocyte ghosts and water. In UV images, blotches are observed due to skin pigmentation, which can be a result of pigmented macules (spots), hyperpigmentation due to sun-damage or conditions such as melasma, or erythematous macules (flat red lesions). Pigmented skin appears as dark patches in UV images as a result of attenuation by melanin in epidermis or induction of collagen cross-links fluorescence in dermis. Polarized imaging can be used to enhance either surface or subsurface skin features [79,95]. Polarized images are obtained by placing two polarizers between the light source and the image-capturing sensor. When the polarizers are perpendicular to each other, a *cross-polarized image* is obtained. In cross-polarized image subsurface features like erythema and lesions are enhanced while the surface features are suppressed due to minimizing of the surface reflection. When the polarizers are parallel to each other, surface features like wrinkles and pores are enhanced. This image is called a *parallel-polarized image*. Polarized imaging has been used to characterize skin features like inflammatory acne, erythema and melanin content [21, 86, 147]. In addition to polarized light and fluorescence imaging, *visual image* is captured in visible light and it captures both surface and subsurface features.

Our appearance-microbiome dataset consists of images of 48 subjects with age varying between 25 and 68. The face of each subject is imaged from left, right and frontal views under five modalities: ultraviolet (UV), blue fluorescence (FLUO), parallel polarization (PPOL), cross polarization (XPOL) and visible light (VISI). Each image is 4032 $\times$ 6048 pixels in size. The most salient attributes are apparent in FLUO and UV images (Figure 3.6), therefore these modalities are used for computational skin modeling. The forehead regions used in our experiments are from the frontal views. Example images of these modalities are shown in Figure 3.2.

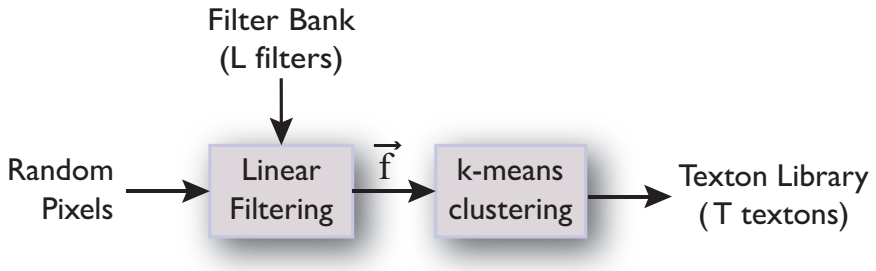


Figure 3.7: **Training phase:Building texton library.** Random sampling of skin images are filtered using a filter bank with L filters. The filter outputs over 5×5 region are clustered using k-means clustering into T textons. For our experiments, we use $L = 48$ and $T = 50$.

3.2.2 Computational Appearance Modeling

In each imaging modality, there are groups of people with perceptually similar skin appearance attributes. In FLUO and UV modalities, we observe the following five skin attributes: *smooth, blotchy, fine hair, and sparse or dense sebum dots* (Figure 3.6). We further categorize the *sebum dots* into red or yellow for the FLUO modality. The skin appearance of subjects can be modeled as a percentage of each of these attributes. This attribute-based approach includes a training phase to obtain a trained neural network (NNET) classifier [136] and an image labeling phase to obtain a skin appearance descriptor.

Training phase: We use two components that are typical in computer vision: texton histograms which is an unsupervised approach, followed by a NNET classifier which is a supervised learning approach. To obtain a texton library (Figure 3.7), a random sampling of skin images are filtered using a filter bank with L filters, resulting in each pixel having an L-dimensional feature vector. Our filter bank is comprised of 48 filters as in [113]. These filters include 36 first and second order derivative of Gaussain filters (6 orientations, 3 scales each), 8 Laplacian of Gaussain filters and 4 Gaussian filters (Appendix A). The filter outputs over 5×5 region are clustered using k-means clustering into T clusters or textons. We empirically choose T=50 for our texton library. Some of the examples belonging to 5 of the 50 clusters are shown in Figure 3.8.

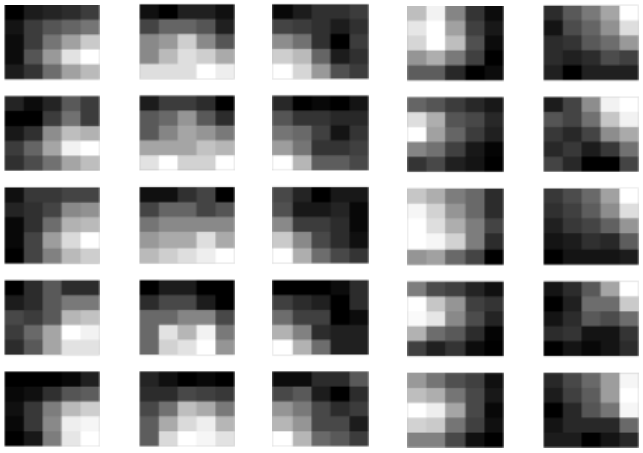


Figure 3.8: **Example patches grouped together to form the texton library.** Similar 5×5 patches are grouped together by clustering their filter responses. Example patches belonging to: (a) Texton 2. (b) Texton 4. (c) Texton 6. (d) Texton 44. (e) Texton 48.

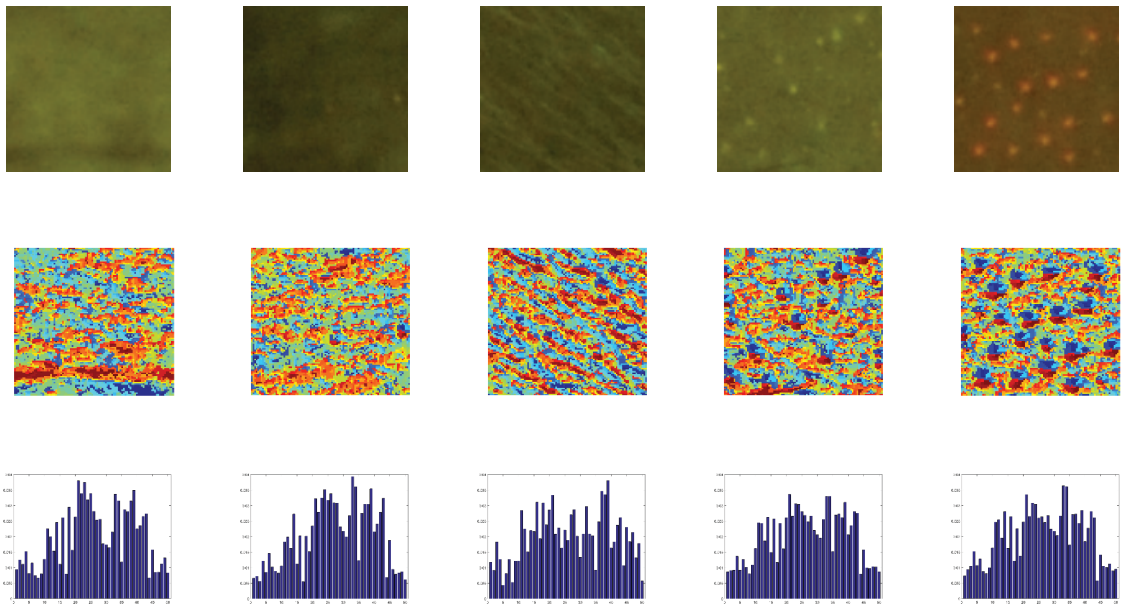


Figure 3.9: **Texton maps and texton histograms.** [Top Row] Example patches in FLUO imaging modality. [Middle Row] First texton map. (c) [Last Row] Soft-weighted texton histogram.

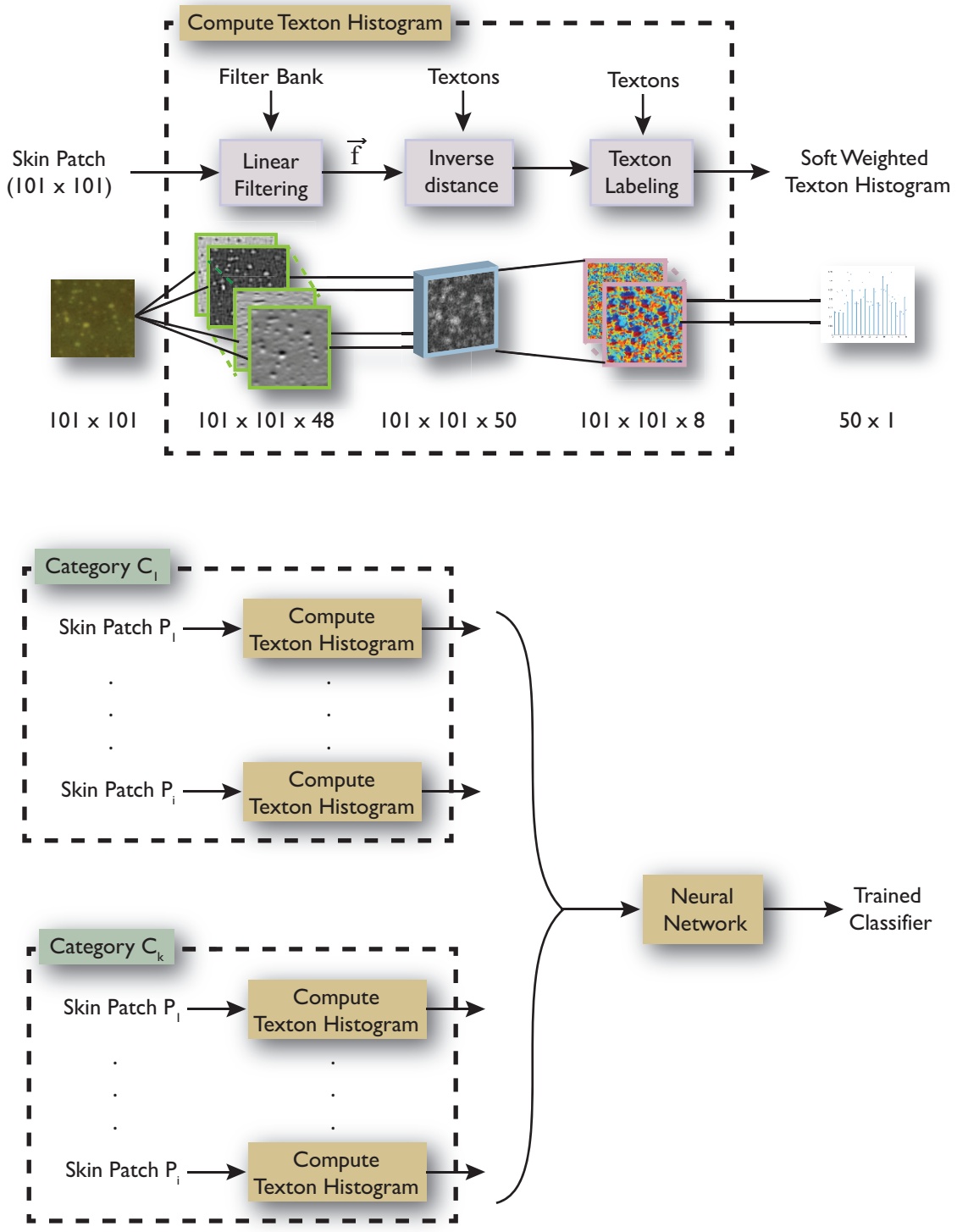


Figure 3.10: **Training phase: Building neural network building (NNET) classifier.** Texton histogram is computed over labeled skin patches of size 101 × 101 and used for training the NNET classifier.

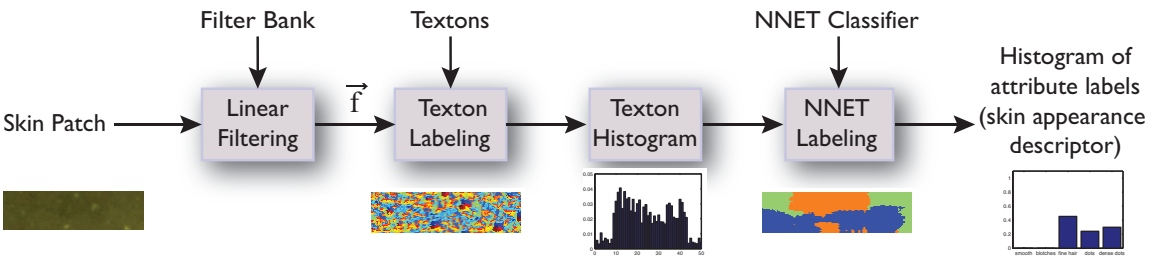


Figure 3.11: **Image Labeling Phase.** The texton histogram of a patch centered at each pixel of a skin image is labeled with one of the attribute categories using a neural networks classifier. The histogram of attribute labels of the entire skin image is its skin appearance descriptor. A texton label characterizes a 5×5 region. An attribute label characterizes a 101×101 region. The histogram of attribute labels describes a larger region (typical size 1000×2000)

A neural network classifier for each modality is trained to classify the skin patches (Figure 3.10). Every patch pixel is assigned the label of its closest texton to obtain a texton map and a texton histogram is computed over each skin patch of size 101×101. We constrict a soft-weighted texton histogram from the 8 nearest textons corresponding to each pixel [169]. These modifications deal with the problem of similar filter responses being assigned to neighboring textons. Figure 3.9 shows some examples of texton maps and texton histograms. The training set is obtained by manually labeling random skin patches with one of the attribute labels described in Figure 3.6. The texton histograms from the labeled skin patches and the patch attribute labels are used for training the neural networks classifier (NNET).

Image labeling phase: is illustrated in Figure 3.11. The term *skin image* refers to the entire extracted forehead region and a histogram of attributes (one attribute per patch) is used to to describe the skin image. For a skin image (typical size 1000×2000), a patch of size 101×101 around each pixel is filtered, labeled with textons and a texton histogram is obtained. Using the texton histogram as input to the trained NNET classifer, the patch corresponding to each pixel is labeled with one of the attributes (for example Figure 3.14). A histogram of attribute labels is then constructed for each skin image, giving a skin appearance descriptor. We merge the attributes labels *sparse sebum dots* and *dense sebum dots* together to form the attribute *sebum dots*. In FLUO modality the color of the dots is an additional attribute that indicates either excitation of porphyrins (red sebum

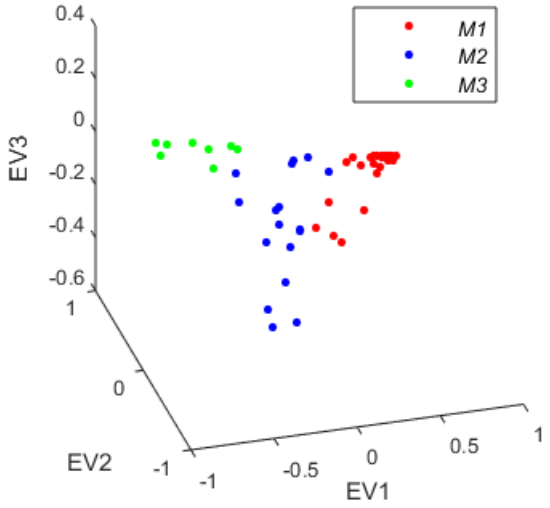
dots) or horn (yellow sebum dots). The dots are detected by finding high gradient pixels that have attribute labels as *sebum dots*. The mean of the normalized red channel for the dot pixels is a measure of dot redness.

Skin appearance of a subject is grouped using the percentage of attributes in each modality. Appearance clusters corresponding to each attribute are defined by specifying a simple threshold on the attribute percentages. For example, when the percentage of pixels in UV labeled as *sebum dots* is high ($\geq 50\%$), that subject is in appearance cluster A_U^D. We define six appearance clusters: A_F^D (percentage of *sebum dot* pixels $\geq 50\%$ and *red color* ≥ 0.76 in FLUO); A_F^B (percentage of *blotchy* pixels $\geq 50\%$ in FLUO); A_F^S (percentage of *smooth* pixels $\geq 50\%$ in FLUO); A_U^D (percentage of *sebum dot* pixels $\geq 50\%$ in UV); A_U^B (percentage of *blotchy* pixels $\geq 50\%$ in UV); A_U^S (percentage of *smooth* pixels $\geq 50\%$ in UV).

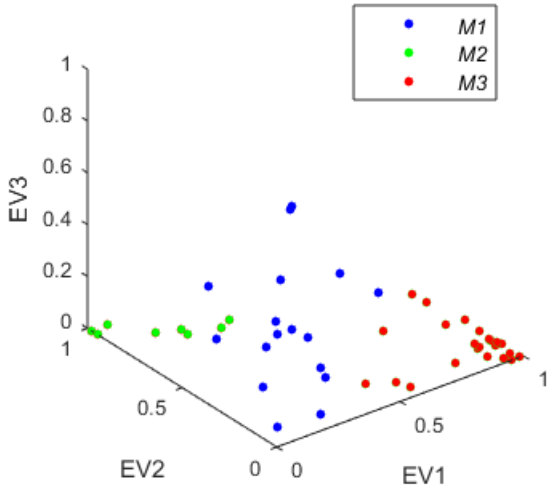
3.2.3 Eigenbiome-Model for Skin Microbiome

Using 16S ribosomal RNA gene sequencing [34, 64, 67], a swab from the forehead of each subject is profiled to obtain the relative abundance of 724 genera. Relative abundance of genus is the concentration (percentage) of each of the 289 genus in a subject's skin microbiome. Out of 724 genera, 289 genra had non-zero relative abundance of genus for all the subjects. Subjects with similar microbiome should group together using clustering techniques. However, clustering in a 289 dimensional space is problematic due to the well-known problem in machine learning referred to as the curse of dimensionality. By projecting this high dimensional data to a lower dimensional subspace, we can obtain meaningful clusters that can be linked to appearance. Additionally, the projection provides a convenient visualization.

Principal component analysis (PCA) is widely used for dimensionality reduction. PCA finds an optimal orthogonal basis set for describing the data such that the variance in the data is maximized. The data can be projected to a lower dimension with eigenvectors which retain the maximum variance. We refer to the microbiome projected to a lower dimensional eigenspace as *eigenbiome*. For the microbiome data, the percentage of variance retained with each eigenvector is analyzed and it is observed that 92.49% variance is retained by



(a)



(b)

Figure 3.12: **Eigenbiome Model.** Microbiome data projected to *eigenspace*: a lower dimensional eigenspace using (a) Principal component analysis (PCA). The eigenbiome vectors for each of the three clusters have both positive and negative components using PCA. (b) Non-negative matrix factorization (NMF). By employing non-negative matrix factorization (NMF), the eigenbiome clusters are constrained to have positive components.

first three eigenvectors. *Thus, three dimensional space is sufficient for this microbiome representation.* Clustering of the eigenbiome is done based on proximity to neighbors by a simple kmeans clustering. The distribution of subjects microbiome in the eigenbiome space suggests clustering with k = 3, i.e., three distinct groups can be visually discriminated. Using three groups, we classify all subjects into one of the three microbiome clusters ($M1$, $M2$ or $M3$) in Figure 3.12(a).

The eigenbiome vectors for each of the three clusters have both positive and negative components using PCA. However, the negative concentrations of relative abundance of genus are not physically realizable. If we employ non-negative matrix factorization (NMF) [111], the eigenbiome clusters are constrained to have positive components (Figure 3.12(b)). This constraint has a very useful physical interpretation. The vector components are positive so that they are *physically realizable* for the relative concentration of genus. Moreover, since NMF favors a sparse solution, the physical interpretation can be enhanced. Sparsity constraints force near zero concentrations to be set to exactly zero. Therefore, the eigenbiome vectors are realizable concentrations of *select* microbes. In this sense the three eigenbiome vectors are distinct microbial communities that contain some microbes, but not others. All subjects are computationally expressed as a mixture of these three dominant communities.

3.2.4 AMCO framework

The secondary microbiome space is high dimensional and clustering using the raw data typically results in clusters where the subject membership is not consistent with the appearance grouping. A metric learning algorithm is used which transforms secondary space to a lower dimension using the class labels transferred from appearance grouping as illustrated in Figure 3.5(c) and 3.13.

Our goal is to learn a transformation in the secondary space such that the data points in the same appearance cluster are near to each other and far from all the data points in the other clusters. The maximally collapsing metric learning (MCML) algorithm [60] aims to project all the points in a class to a single point and in this ideal scenario the conditional

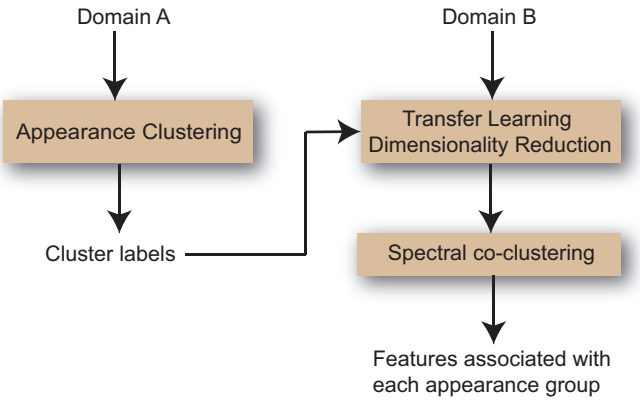


Figure 3.13: **Appearance-driven multiview co-clustering (AMCO) framework:** Proposed methodology to link groups in appearance space A to features in the secondary space B

probability for each data point i over all other points j is given by:

$$p_0(i \mid j) = \begin{cases} 1 & c_i = c_j \\ 0 & c_i \neq c_j \end{cases}, \tag{3.1}$$

where c_i are the class labels from appearance clustering transferred to the secondary space.

To project the data points in each cluster to a single point, the distance among the points in the transformed space must be minimized. Consider the Mahalanobis distance between two points given by:

$$d_{ij}^A = (x_i - x_j)^T A (x_i - x_j), \tag{3.2}$$

where x_i, x_j are data points in the secondary space, A is a positive semi-definitive matrix.

Using this distance metric, the conditional probability which approximates Equation 3.1 is given by [60, 62]:

$$p_A(j \mid i) = \frac{e^{-d_{ij}^A}}{\sum e^{-d_{ij}^A}}. \tag{3.3}$$

For Equations 3.1 and 3.3 to be approximately same, the matrix A can be found by minimizing the KL divergence of both the distributions:

$$\begin{aligned} \underset{A}{\text{minimize}} \quad & \sum_i KL\left[p_0(j \mid i) \mid p_A(j \mid i)\right] \\ \text{subject to} \quad & A \in PSD. \end{aligned} \tag{3.4}$$

This optimization problem is convex and can be solved for matrix A using a standard solver or an iterative approach. Matrix A transforms the raw data to a new metric space. Further, to project the data to a k-dimensional subspace, A can be decomposed into eigenvalues and eigenvectors and A^k can be constructed from the eigenvectors corresponding to k highest eigenvalues.

Spectral co-clustering extends spectral clustering that uses the eigenvectors of an affinity matrix to find clusters. The data points to be clustered based on similarity measure can be represented as a similarity graph. The data points are the vertices of the graph and any two vertices are connected by an edge if they are similar. Each edge has a weight associated with it, which is the measure of similarity between the vertices connected by that edge. Partitioning the graph based on the weights associated with each edge clusters the similar data points. In [35], a bipartite graph is used to find co-clusters, i.e., simultaneously cluster rows (subjects) and columns (features). A bipartite graph is a graph whose vertices can be divided into two disjoined sets such that the vertices within a set are not connected and the edges connect a vertex from one set (subjects) to a vertex in another set (features). The subjects and features are linked to each other using edges.

For a fully-connected graph, the similarity or affinity matrix is computed using Gaussian kernel as

$$A = e^{\frac{-\|x_i - x_j\|^2}{2\sigma^2}} \tag{3.5}$$

where $x_i \in R^d, i = 1...N$ are the d-dimensional data points. The symmetric normalized graph Laplacian of the affinity matrix, which has a block diagonal form is computed as:

$$L = D^{-0.5} A D^{-0.5} \tag{3.6}$$

where D is the diagonal matrix with $D(i,i) = \sum_i A_{ii}$.

The normalized graph Laplacian matrix can be obtained by computing the eigenvectors

of the generalized problem $Lu = \lambda Du$ as in [167] or by normalizing the rows of matrix U to norm 1 as in [140]. Singular value decomposition of matrix L gives left and right eigenevectors corresponding to data points and features, respectively. By clustering these eigenvectors, similar subjects *and* their corresponding features are grouped in clusters.

3.3 Experiments and Results

To obtain the skin appearance descriptors, a neural network classifier for each modality is trained to classify the skin patches as explained in Section 3.2.2. Training and test sets were obtained from different subjects. Each skin patch of size 101×101 is manually assigned to one of the five categories of skin attribute patches: smooth (SM), blotchy (BL), fine hair (FH), sparse sebum dots(SD) and dense sebum dots (DD). Texton histogram of each skin patch is used as a feature vector to train the classifier. The number of training and test samples for each category are listed in Table 3.2. Table 3.3 shows the effect of using soft weighting on the classifier accuracy. When both the methods are integrated in the approach, the classifier accuracy improves for both FLUO and UV modalities.

Attribute	Training	Test
SM	200	200
BL	400	400
FH	200	0
SD	400	400
DD	400	400

Table 3.2: Training and test sample size in each category. SM: smooth; BL: blotchy; FH: fine hair; SD: sparse dots; DD: dense dots.

	FLUO		UV	
	Training	Test	Training	Test
Without Soft Weighting	92.9	89.5	93.1	84.4
With soft weighting	**96.5**	**95.4**	**96.8**	**88.5**

Table 3.3: NNET accuracy with soft weighting. Note that soft weighting improves test accuracy for both FLUO and UV datasets.

Using the computational appearance modeling discussed in Section 3.2.2, the forehead of a subject in each modality is labeled using the trained NNET classifier as illustrated in Figure 3.14. The histogram of attributes is the skin appearance descriptor. The appearance

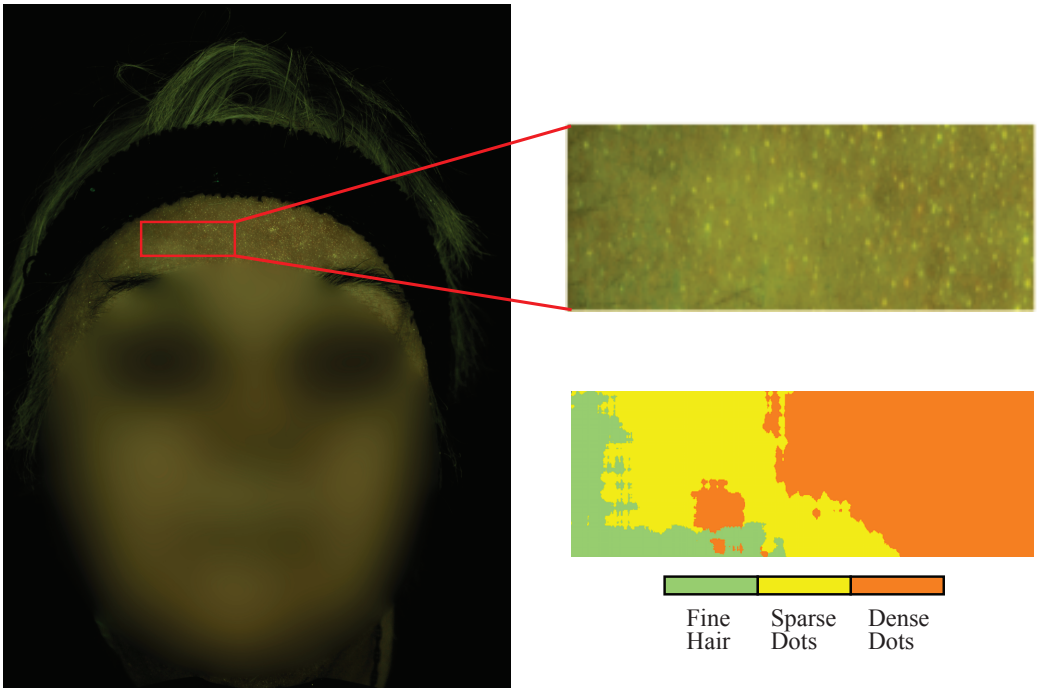


Figure 3.14: **Image labeling using NNET classifier.** The forehead skin image from the frontal view of the subject in FLUO modality has been labeled using NNET classifier. Face is blurred to preserve the privacy of the subject.

clusters defined by specifying thresholds of attributes. The six appearance clusters, three each for FLUO and UV modalities, are listed in Table 3.4. These appearance clusters are of interest because our results indicate a clear microbiome association. There were not many subjects in the dominant fine-hair category, so a connection to the microbiome could not be made and the category is omitted from Table 3.4.

Using the eigenbiome model in Section 3.2.3 each subject is projected to a three dimensional space using PCA and assigned to one of the three microbiome clusters ($M1$, $M2$ or $M3$). Figures 3.15 shows the overlap of microbiome cluster $M1$ and appearance cluster A_F^D (high concentration of sebum dots with a red color in FLUO modality) shows that this appearance cluster is linked to microbiome cluster $M1$. Using non-negative matrix factorization (NMF) for projecting the microbiome to a lower dimensional space (Figure 3.16), the eigenbiome clusters are constrained to have positive components so that they are physically realizable for the relative concentration of genus. The subjects grouped together using the projected microbiome data by NMF are same as the subjects in groups using PCA.

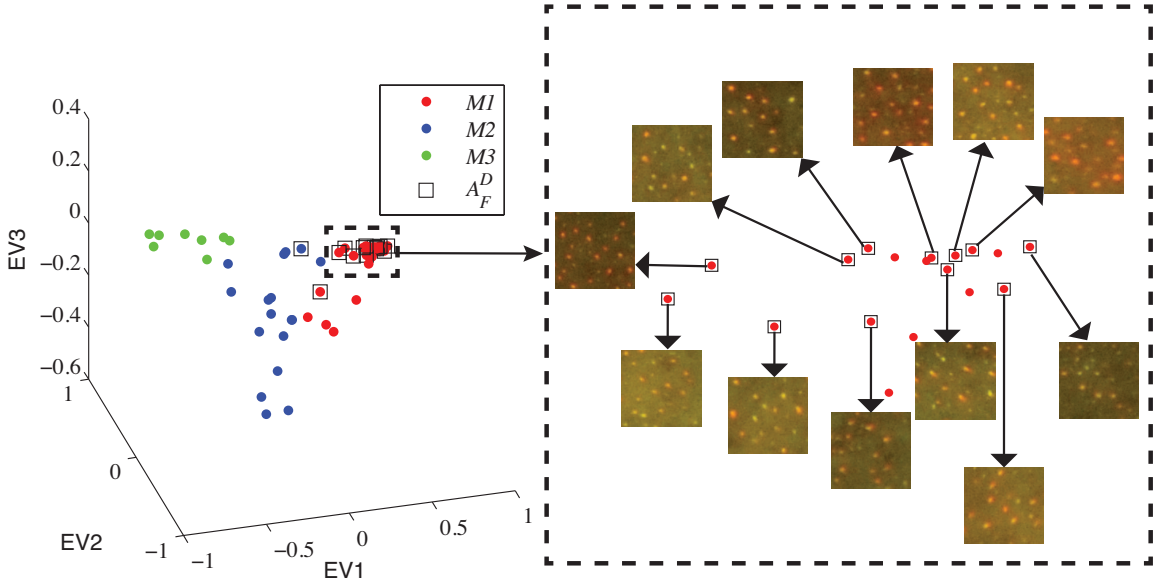


Figure 3.15: **Clusters in eigenbiome have been linked to appearance clusters.** (Left) Microbiome is projected to a three dimensional space using PCA and three clusters $(M1(red), M2(blue), M3(green))$ are found using kmeans clustering. The rectangular markings show the appearance cluster A_F^D (high concentration of sebum dots with red color threshold$\geq$0.76 in FLUO modality). Overlap of microbiome cluster $M1$ and appearance cluster A_F^D shows that this appearance cluster is linked to microbiome cluster $M1$. (Right) Example patches are shown for the subjects that are in the overlap of appearance cluster A_F^D and microbiome cluster $M1$. The conditional probabilities such as $\mathrm{P}(M1 \mid A_F^D)$ are given in Table 3.4. As expected $\mathrm{P}(M1 \mid A_F^D)$ is high showing that appearance is predictable of the microbiome cluster.

Appearance Cluster A	n(A)	P($M1 \mid A$)	P($M2 \mid A$)	P($M3 \mid A$)	P($A \mid M1$)	P($A \mid M2$)	P($A \mid M3$)
A_F^D: FLUO modality Sebum dot $\geq 50\%$ Color ≥ 0.76	14	**0.93**	0.07	0	0.57	0.06	0
A_F^B: FLUO modality Blotchy $\geq 50\%$	5	0	0.6	0.4	0	0.18	0.25
A_F^S: FLUO modality Smooth $\geq 50\%$	7	0	**0.85**	0.14	0	0.35	0.13
A_U^D: ULVI modality Sebum dot $\geq 50\%$	29	0.66	0.17	0.17	0.83	0.29	0.63
A_U^B: ULVI modality Blotchy $\geq 50\%$	12	0.25	0.58	0.17	0.13	0.41	0.25
A_U^S: ULVI modality Smooth $\geq 50\%$	6	0	**0.83**	0.17	0	0.29	0.13

Table 3.4: **Conditional probabilities** for microbiome clusters $M1, M2, M3$ and appearance clusters based on the following appearance attributes: *Dots* - high concentration of sebum dots pixels with red color above threshold, *Blotchy* - high concentration of blotchy pixels and *Smooth* - high concentration of smooth pixels. Observe that the probability of microbiome $M1$ conditioned on appearance cluster A_F^D in FLUO modality with a high concentration of sebum dots and redness=0.76 is 0.93, indicating a 93% chance of a subject being in microbiome cluster $M1$ given this appearance cluster. Similarly observe that the probability of microbiome $M2$ given a high concentration of smooth pixels in FLUO and UV is high (0.85 and 0.83 given A_F^S and A_U^S, respectively.)

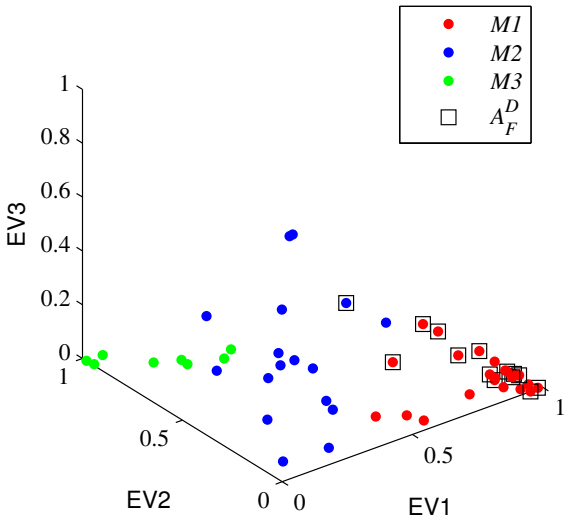


Figure 3.16: **Non-negative matrix factorization (NMF) for microbiome projection.** Using NMF for projecting the microbiome to a lower dimensional space, the eigenbiome clusters are constrained to have positive components so that they are physically realizable for the relative concentration of genus. Overlap of microbiome cluster $M1$ and appearance cluster A_F^D(high concentration of sebum dots with red color threshold$\geq$0.76 in FLUO modality) shows that this appearance cluster is linked to microbiome cluster $M1$.

Table 3.4 shows the conditional probability of each of three microbiome clusters conditioned on the individual appearance cluster. High conditional probabilities indicate a high likelihood of the microbiome cluster when the subject exhibits the particular appearance attribute. Observe that in three distinct cases the conditional probability is high: 1) A_F^D: *sebum dots with a red color above the indicated threshold in FLUO* (predictive of microbiome cluster $M1$ with $P(M1 \mid A_F^D)$=0.93); 2) A_F^S: *smooth in FLUO* (predictive of microbiome cluster $M2$ with $P(M2 \mid A_F^S)$=0.85); and 3) A_U^S: *smooth in UV* (predictive of microbiome cluster $M2$ with $P(M2 \mid A_F^S)$=0.83). For the appearance cluster A_F^D) the conditional probability increases as redness of dots increases but the number of samples in the appearance cluster decreases (see Figure 3.18). Our results reveal a strong link between appearance clusters (captured instantaneously with camera) and microbiome clusters (from time-consuming sequencing).

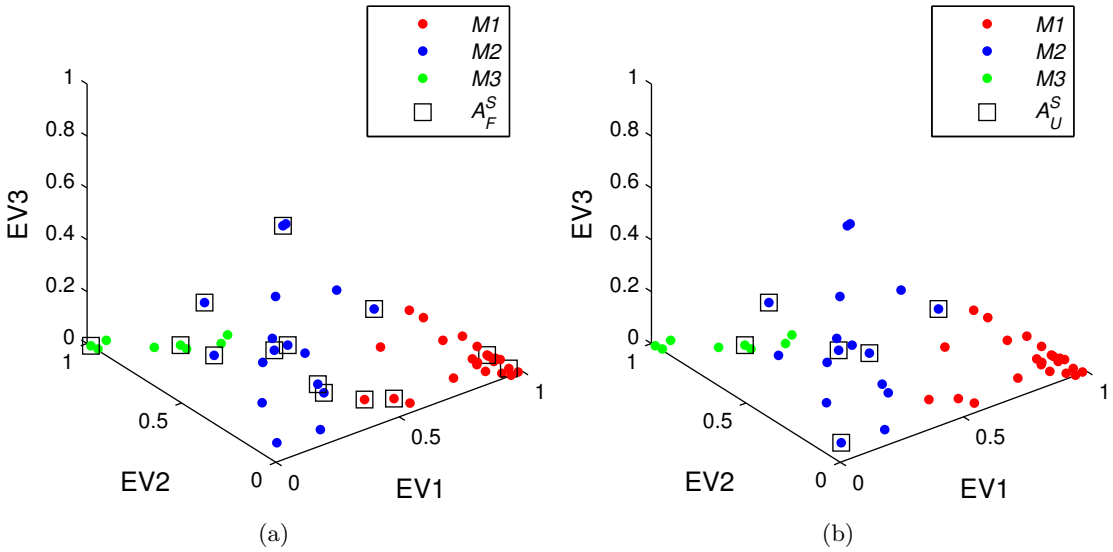
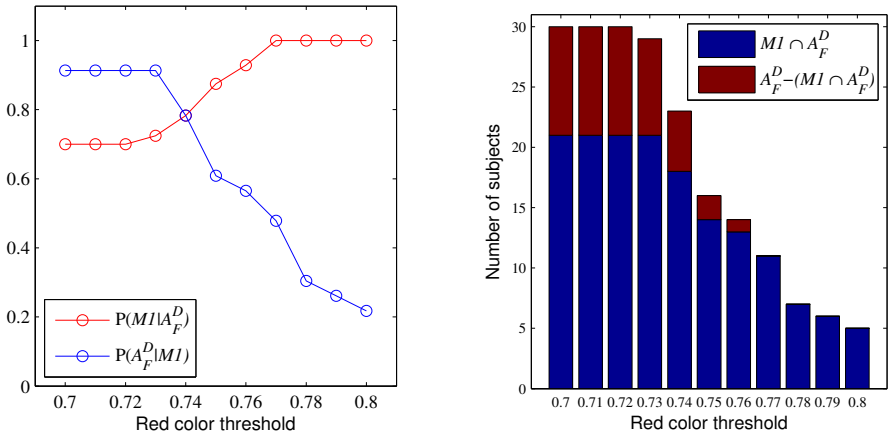


Figure 3.17: **Clusters in eigenbiome have been linked to appearance clusters.** Microbiome is projected to a three dimensional space and three clusters $(M1, M2, M3)$ are found using kmeans clustering. (a) Overlap of the appearance cluster A_F^S (high concentration of smooth pixels in FLUO modality) shows that this appearance cluster is linked to microbiome cluster $M2$. (b) Overlap of the appearance cluster A_U^S (high concentration of smooth pixels in UV modality) shows that this appearance cluster is linked to microbiome cluster $M2$.



(a) (b)

Figure 3.18: **Conditional probability as a function of redness of dots.** (a) Conditional probability for linking of microbiome from appearance $P(M1 \mid A_F^D)$ and appearance from microbiome $P(A_F^D \mid M1)$ as a function of redness of dots. For varying threshold of sebum dot redness, appearance cluster A_F^D has subjects with high concentration of sebum dots ($\geq 50\%$) with indicated red color threshold in FLUO. (b) Number of subjects in clusters A_F^D and $M1$ ($M1 \cap A_F^D$) as a function of redness of dots. As the threshold increases, the conditional probability of a subject to be in microbiome cluster $M1$ given it is in appearance cluster A_F^D increases whereas the number of subjects in appearance cluster A_F^D decreases.

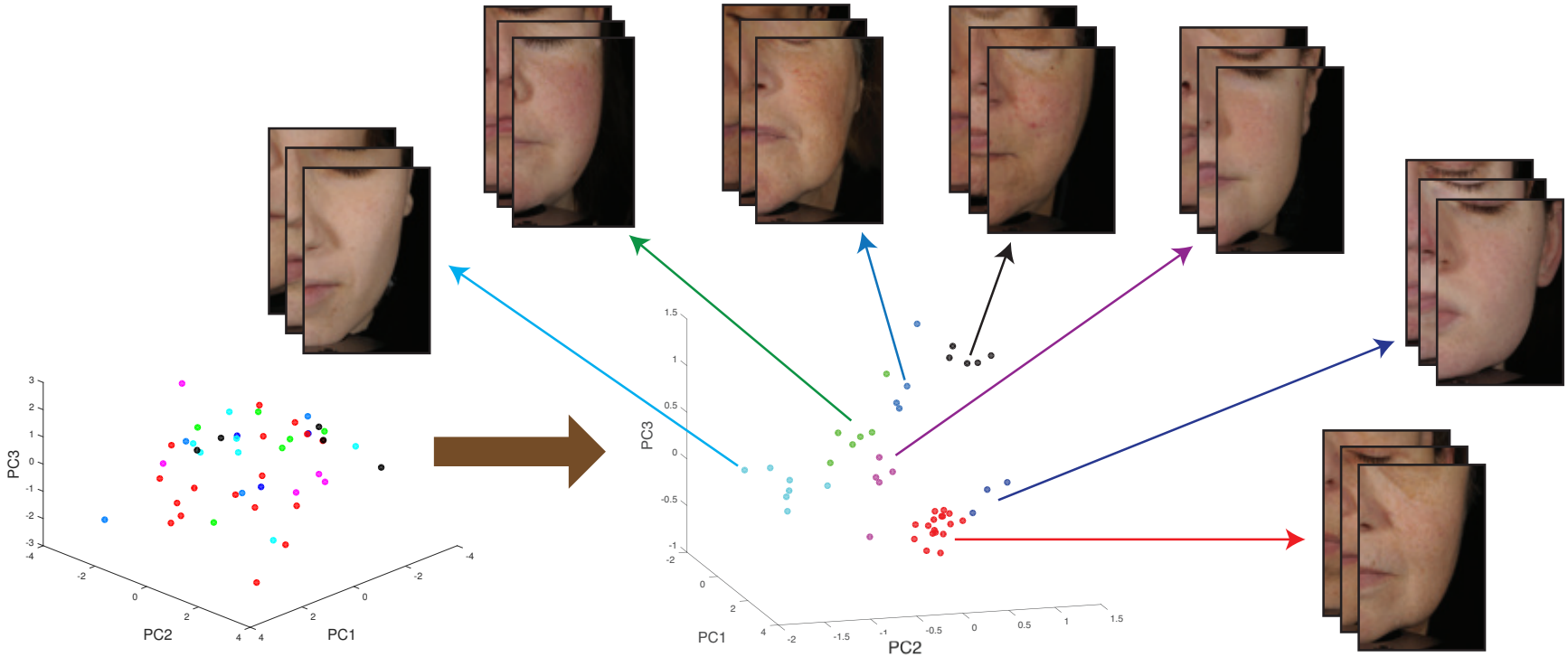
After establishing the link between skin appearance and microbiome, we present the results for AMCO framework. In addition to the appearance-microbiome datasets in FLUO and UV imaging modalities, we also include the results for following datasets:

1. *Appearance(XPOL)-microbiome* dataset consists of 47 subjects. The cheek regions of the subjects are clustered into 7 classes according to the redness of their skin. The secondary space is the 365-dimensional cheek microbiome for each subject.

2. *Animals with Attributes* [106] consists of 50 types of animals with 85 attributes per category. In our experiments, we manually group the given animals into 12 classes.

All the datasets are binary, with value 1 indicating the presence of an attribute/microbiome and 0 indicating its absence.

Figure 3.19 shows the data in secondary (microbiome) space for appearance (XPOL)-microbiome dataset. For visualization, the data are projected to three dimensions using principal components analysis. Notice that the data points in appearance clusters do not group together in microbiome space (each color represents an appearance cluster). By using AMCO, transferring the cluster labels in appearance space to microbiome space and projecting the data to a lower dimensional metric based subspace, the data points form meaningful clusters. Similarly, we observe the meaningful clusters in other datasets after transfer-learning metric based projection as shown in Figure 3.20.



(a)

Figure 3.19: **AMCO: appearance(XPOL)-microbiome dataset.** [Left] The data in secondary (microbiome) space (projected to 3D for visualization). The appearance clusters in XPOL modality do not group together in secondary space (each color represents an appearance cluster). [Right] The clusters are visually distinguishable after AMCO. Also see Figure 3.20 and Appendices B and C for results on appearance(FLUO, UV)-microbiome and animals with attributes datasets.

We compare the performance of AMCO with clustering performance on raw data as well as other dimensionality reduction algorithms which do not use labels transfered from appearance clustering. These methods include principal component analysis (PCA), non-negative matrix factorization (NMF) [110] and classical multidimensional scaling (MDS) [189]. In addition, we also use large-margin nearest neighbor (LMNN) [192] method by transferring the appearance cluster labels to secondary domain. Normalized mutual information (NMI) is used to evaluate the clustering quality of data points between true clustering (from appearance domain) and obtained clustering (using AMCO). NMI varies between 0 and 1, with larger value indicating a consistency between the appearance clusters and secondary space clusters. Table 3.5 shows the mean NMI and standard deviation averaged over 20 experiments. The NMI on animal improves from 0.6896 to 0.9107. For all the three microbiome datasets (FLUO, UV, XPOL), NMI increases considerably from 0.1551, 0.2097 and 0.1436 to 0.9587, 0.9527 and 0.8373, respectively. The standard deviation for all the datasets was less than 0.04.

	raw	PCA	NMF	MDS	LMNN	**AMCO**
Animals with Attributes	0.6896	0.7005	0.6093	0.7022	0.5650	**0.9107**
Appearance (FLUO) - Microbiome	0.1551	0.1016	0.1184	0.1096	0.0444	**0.9587**
Appearance (UV) - Microbiome	0.2097	0.0628	0.1263	0.0668	0.1078	**0.9527**
Appearance (XPOL) - Microbiome	0.1436	0.2274	0.1790	0.2296	0.2158	**0.8373**

Table 3.5: **Normalized Mutual Information (NMI)** to evaluate the clustering quality of data points between true clustering from appearance domain and obtained clustering using AMCO. Larger values of NMI indicate a consistency between the appearance clusters and secondary space clusters. Notice the significantly larger NMI values for the AMCO approach.

In the animals with attributes dataset, when co-clustering on the raw attribute dataset (not using AMCO), some of the resulting groups are: 1) mole, mouse, hamster, rabbit, squirrel, skunk (associated features: small, weak, buckteeth, tunnels, hops); 2) rat, weasel, siamese cat, racoon, dalmatian, collie, persian cat, chihuahua, polar bear, beaver, otter (associated features: forest, nestspot, forager, scavenger); 3) chimpanzee, giant panda, gorilla, spider monkey (associated features: fierce, smart, hunter, meat, muscle, meat-teeth, lean, pads, mountain). Note that this grouping doesnt make sense in terms of its subject membership. Also, each feature can be associated with only one cluster. After AMCO, we

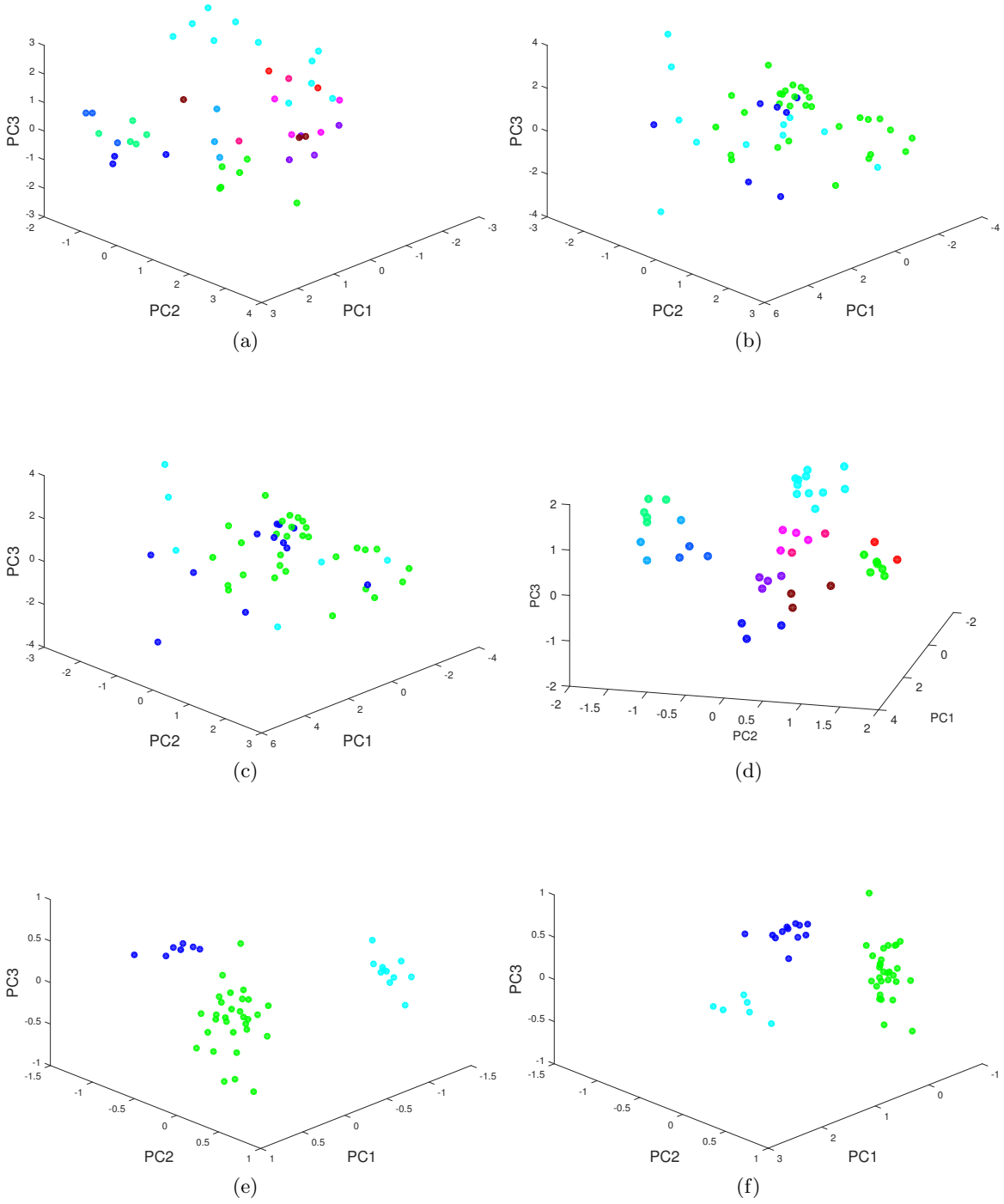


Figure 3.20: **AMCO: animals with attributes and appearance-microbiome datasets.** [Top row] The data in secondary (attribute/microbiome) space (projected to 3D for visualization). The attributes/appearance clusters in do not group together in secondary space (each color represents an appearance cluster). [Bottom row] The clusters are visually distinguishable after transferring the cluster labels to secondary space and projecting the data to a lower dimensional metric based subspace. (a),(d) animals with attributes dataset. (b),(e) appearance(FLUO)-Microbiome dataset. (c),(f) appearance(UV)-Microbiome dataset.

observe the following interesting clusters: 1) mole, mouse, hamster, rabbit, squirrel, skunk, rat, weasel, otter, beaver 2) spider monkey, gorilla, chimpanzee. 3) persian cat, siamese cat. 4) polar bear, giant panda, grizzly bear.

We include datasets with attributes as a secondary space, because the grouping can be directly evaluated to determine if the AMCO method in creating more meaningful groups in the attribute space. However, the true utility of our approach is for finding a latent secondary space grouping that can only be evaluated by its consistency with the appearance group. For our skin dataset, we are exploring a yet undiscovered link between appearance and skin microbiome, demonstrating causative appearance analysis. When a group of microbiome features is associated with a particular appearance group, that microbiome community may be a potential cause of the appearance [92]. This relationship between appearance and microbiome was first explored in [48] but with a rudimentary "good", "bad" description of skin appearance. This microbiome/appearance association represents an important example of a new frontier in computer vision where visual modeling drives discovery of how appearance depends on biological, manufacturing, weather, agricultural or other process spaces.

3.4 Conclusions

In this chapter, we present an attribute-based appearance model using texton-analysis of blue fluorescence and ultraviolet imaging modalities. Using 48 subjects, we link appearance to the eigenbiome, the low dimensional representation of a subject's skin microbiome. The eigenbiome model using non-negative matrix factorization represents physically realizable concentrations of microbes. The intersection of the appearance and eigenbiome clusters reveals three interesting cases where the probability of a subject belonging to a microbiome cluster conditioned on appearance is high. The sequencing of microbiome takes several days but computational appearance is obtained in seconds. The established link to microbiome clusters provides biological information with photographic imaging.

Further, we demonstrate appearance-driven multiview co-clustering framework that extracts potentially causative clusters by tuning the secondary space so that the grouping

matches appearance grouping. We demonstrate results on an appearance-microbiome data set where the latent grouping in microbiome space is revealed and associated with appearance. Additionally, we demonstrate results on animal-attribute dataset where the grouping can be easily checked to verify the results.

Chapter 4

Classification of Microscopic Skin Images

Reflectance Confocal Microscopy (RCM) is a non-invasive technology that is used to diagnose and study skin cancer, skin aging, pigmentation disorders and skin barrier function by measuring thickness of different skin layers [78, 181]. The effect of skin treatments that modify the proliferation processes within the skin can also be measured through thickness changes. RCM images individual skin layers at different depths by the optical sectioning property of the composite lenses and apertures. A laser is used as the monochromatic light source and tissue penetration is wavelength dependent. As the wavelength increases, the penetration depth also increases; however, higher wavelengths result in tissue damage. A typical penetration depth is 200-250 μm for a wavelength of 800-1024 nm [161]. The light passes through a beam splitter, a focusing lens and is focused on a small tissue spot of skin (few microns) [11]. Each skin layer has different cellular structures causing variation in light reflection, refraction, absorption and transmission. The reflected light passes through an objective lens, a pinhole filter and is imaged at the photo-detector. Captured images of each skin layer have an observable image texture that is unique for each skin layer as shown in Figure 4.1.

RCM captures the cellular details at a spatial resolution that is comparable to histopathology, which is an invasive, painful and time consuming procedure. Using RCM, a series of images are acquired at the same position, from epidermis through upper dermis and are collectively called a stack. The epidermis is divided into four sub-layers: stratum corneum (SC), stratum granulosum (SG), stratum spinosum (SS) and stratum basale (SB) ((Figure 4.1(b)-(e)). Portions of papillary dermis are also imaged (Figure 4.1(f)). In addition, certain images are categorized as outside the dermis (OD) (Figure 4.1(a)).

Traditionally, these high resolution stack images are labeled based on visual observation

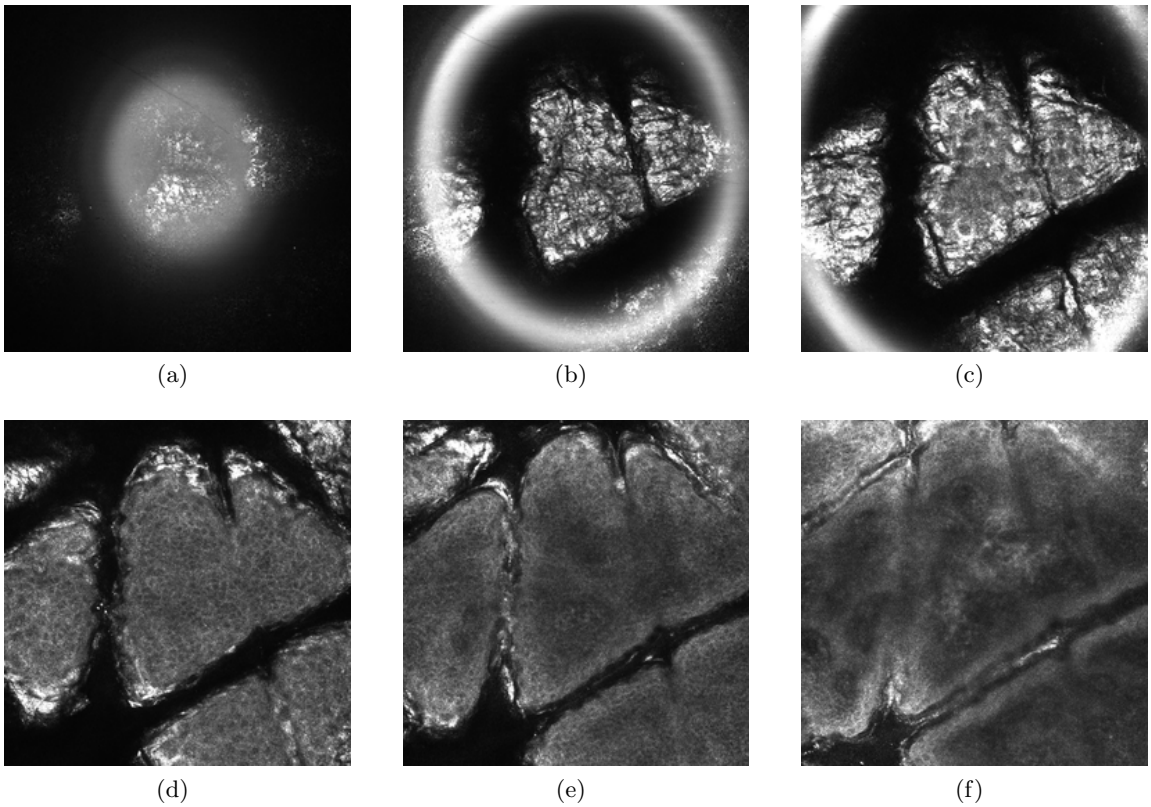
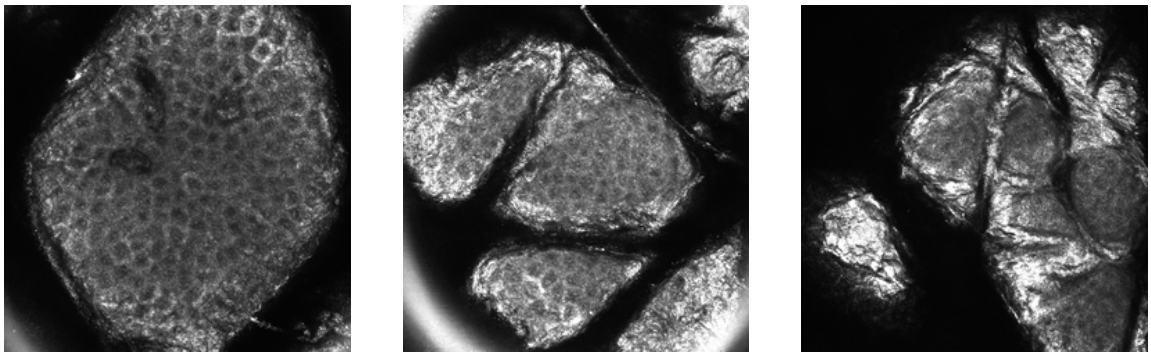


Figure 4.1: **Skin layers in an RCM image stack.** (a) Outside epidermis (OE). (b) stratum corneum (SC). (c) stratum granulosum (SG). (d) stratum spinosum (SS). (e) stratum basale (SB). (f) portions of the papillary dermis (PD).

by clinical experts. This manual labeling requires significant time and also results in labeling variability depending on the expertise of the clinical grader. We propose an automated method based on detecting skin features and training a mutilayer neural network. Our method is a hybrid of classic methods in texture recognition called *texton-based recognition* and recent methods of *deep learning*. Our results indicate that this hybrid approach gives higher recognition rates than CNN deep learning methods and other recent texture recognition methods. Moreover, in situations where moderate size image sets are available (thousands as opposed to millions) the results are significantly better (higher accuracy) and faster than the competing approaches.

A key component in CNN research is the ability to obtain optimal filters with automatically derived features. However, the lowest level of these networks look very similar to traditional filter banks. The CNN methods were developed for the task of general image



(a) (b) (c)

Figure 4.2: **Intra-class variation.** Images of the same class from different RCM stacks have similar texture but the overall structure can change significantly and therefore the recognition problem is quite challenging.

recognition [88,100]. For RCM labeling, the task is texture recognition where the number of classes is small and the image variation is subtle. Texture recognition methods have traditionally relied on multiscale, multi-orientation, gradient-like features called textons that are paralleled in the human visual system [20, 22, 23, 113, 164, 184]. Texton-based features have the advantage that they are computed from fixed weight filters and need not be trained. More recently texture recognition has been addressed using deep learning and CNN [17]. We develop a hybrid approach that uses texton-based features as input to train a deep neural network in order to gain the advantages of both traditional and recent texture recognition frameworks.

Our dataset comprises of 1500 skin images from 15 RCM stacks, each image belongs to one of the skin layers as shown in Figure 4.1 where labels were obtained by clinical skin experts. Figure 4.2 shows images of the same class but from different RCM stacks (one stack corresponds to the images obtained by varying the depth through one skin sample). Notice that the texture within each class is similar, but the overall structure can change significantly and therefore the recognition problem is quite challenging.

We demonstrate that our hybrid deep learning approach performs with a test accuracy of 81.73%. In most stacks, mislabeling is observed between adjoining layers as the images transition from one skin layer to another. These transitions may be ambiguous during manual labeling as well.

4.1 Related Work

Computational analysis of RCM skin images has been used for automatic detection of tumors [94], detecting malignant features for superficial spreading melanoma [51] and skin aging assessment [152]. Texture of RCM images has also been analyzed to identify melanocytic skin lessions at dermal-epidermal junction [98] using Speeded Up Robust Features (SURF) to capture the texture of localized features and use Support Vector Machine (SVM) classifier to distinguish between them.

Automated methods to categorize RCM stack images into skin layers has been recently proposed in [71, 72, 103, 174]. Delineation of dermal-epidermal junction is presented in [103] but is limited to categorizing dermis and epidermis skin layers based on the difference in their contrast. The sub-layers in epidermis are not identified. To categorize each image in the RCM stack as one of the skin layers, a fully unsupervised texton-based approach is presented in [174]. A texton library is created by using a filter bank and then projecting the filter response to a lower dimensional subspace using principal component analysis. The texton histogram of images are projected to a lower dimensional subspace and clustered into five skin layers using k-means clustering results. A high correlation is reported between the ground truth and obtained labels. However, only three stacks are used in these experiments for evaluation.

Bag-of-features representation of images has been used in computer vision for tasks such as object or scene recognition [47, 194]. Local interest features of iconic patches are used to build a dictionary and each image is represented by frequency of each visual word in the dictionary. In [71], the authors apply a bag-of-features approach and use logistic regression classifier. A representative dictionary is built by combining hierarchical and k-means clustering for normalized 7×7 patches. In [72], conditional random fields are followed by structured SVM as a classifier instead of logistic regression and shows improved performance. For our approach, we instead use a deep-learning framework that is generally more powerful in discrimination when compared to the SVM classifier. We also empirically demonstrate that the filter-based texton histogram is a better feature to classify RCM skin images compared to patch-based features.

Perceptual attributes have been used in describable texture database where each image is assigned several attributes based on perception, inspired by the human vision [17]. In [92] perceptual attributes are used to categorize macroscopic skin textures. We extend this method by using the attribute histograms to train another classifier which can be used to label RCM images. However, we observe that even though this approach provides pixel level attribute labels, the test image accuracy using attribute histograms for training a neural network on the RCM data is relatively low (74.94%). Further details about this method are given in Secion 4.2.3.

Convolutional Neural Networks (CNNs) have been used successfully for several computer vision tasks such as image classification, video analysis and object recognition [17, 39, 88]. CNNs learn a hierarchy of features for classification automatically from a large set of input images (in millions). We demonstrate that a CNN trained on a moderate size dataset results in low test accuracy. The CNN architecture we use is presented in Section 4.2.4. In RCM skin images, the differences in the textures of various categories are very subtle (Figures 4.1) and the datasets are relatively small, which makes skin classification a challenging problem.

4.2 Methods

4.2.1 Imaging

The acquisition of RCM stacks is done using a Vivascope 1500 (Lucid Technologies, Rochester, NY, USA) using 785 nm laser illumination. The image stack was collected up to the depth of 100μm, at a step size of 1μm. The dimensions of the images are 1000×1000 pixels.We collected a dataset consists of 15 stacks, with 100 images in each stack. Each image is labeled by a human dermatology expert as one of the following skin layer category as illustrated in Figure 4.1: Outside epidermis (OE), stratum corneum (SC), stratum granulosum (SG), (d) stratum spinosum (SS), (e) stratum basale (SB), (f) portions of the papillary dermis (PD).

4.2.2 Hybrid Deep Learning

Our hybrid deep learning approach combines the unsupervised texton-based approach with supervised deep neural networks. It consists of the following layers as shown in Figure 4.3:

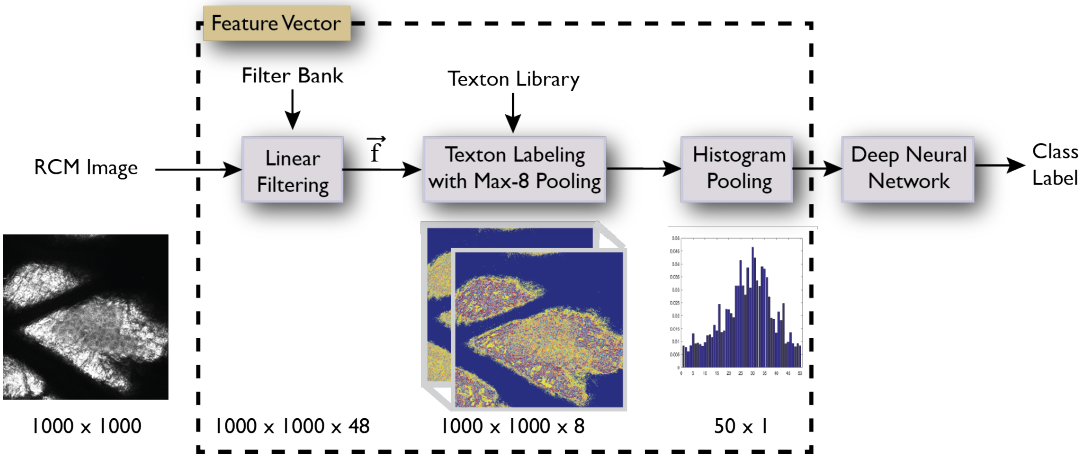


Figure 4.3: **Hybrid Deep Learning.** Texton-based feature vectors are obtained by using a pre-built texton library. Patches (5×5) centered at each pixel of the RCM image are labeled with 8 closest textons in the texton library. These texton labels are pooled to obtain a texton histogram which is used as input to train a deep neural network with multiple layers.

1. *Convolution layer:* We use a fixed-weight filter bank with 48 filters. These filters include 36 first and second order derivative of Gaussain filters (6 orientations, 3 scales each), 8 Laplacian of Gaussain filters and 4 Gaussian filters ([113]). Each pixel is filtered over a 5×5 region and represented by a 48-dimensional vector.

2. *Texton labeling with Max-8 Pooling:* Patches (5×5) centered at each pixel of the skin image are labeled using a pre-built texton library. The texton library is obtained by clustering the 48-dimensional filter outputs of a random sampling of skin images over 5×5 region into T clusters. For our experiments we use k-means clustering with $T = 50$ clusters. The texton library needs to be built only once. Using the texton library, 48-dimensional filtered output of each pixel is mapped to its 8 closest textons from the texton library. Each pixel is associated with 8 nearest cluster centers resulting in 8 texton maps for each RCM image. We associate each pixel to 8 clusters instead of 1 so that the effect of the similar filter responses being assigned to neighboring textons is nullified in next layers [169].

3. *Histogram Pooling:* The textons labels from the 8 texton maps are pooled together by weighing them their distance. For each pixel, p, the histogram bin h corresponding

to its texton labels is updated as:

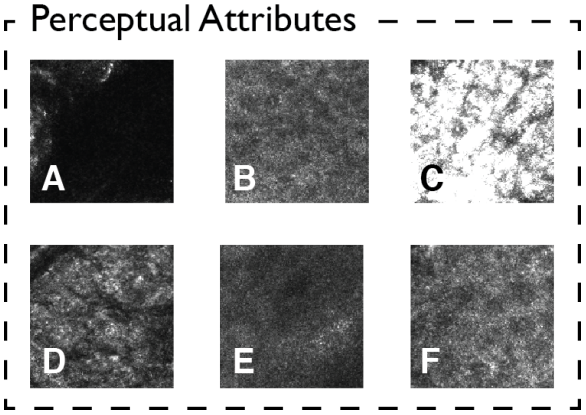
$$h(t) = h(t) + \left[1 - \frac{d(i)}{\sum_{i=1}^{8} d(i)}\right], \tag{4.1}$$

where $d(i)$ is the distance of the i^{th} closest texton to the filtered output at pixel p, t is the i^{th} closest centroid in the texton library for the filtered output at pixel p. Texton labels corresponding to dark pixels are ignored and improve the classifier performance.

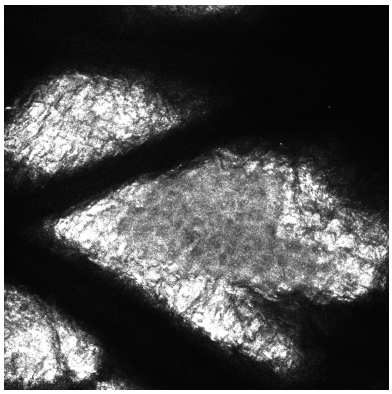
4. *Deep Neural Network:* We use a feed-forward deep neural network, which consists of an input layer, two hidden layers and an output layer [135]. The input and output layers have 50 and 6 neurons, respectively. The two hidden layers have 40 and 10 neurons, respectively. Tan-sigmoid function is used for activation of the neurons. The network parameters were tuned empirically using the training data. Adding more hidden layers or neurons in each hidden layer does not improve the performance for this dataset. An advantage of using the deep neural network is that it returns probability estimates for all the class instead of a single class label, which can also be easily converted to class labels using a linear transfer function at the output layer. The texton histogram of training data is used to train the network, which can then be used to categorize the test skin images into skin layers.

4.2.3 Attribute-based approach

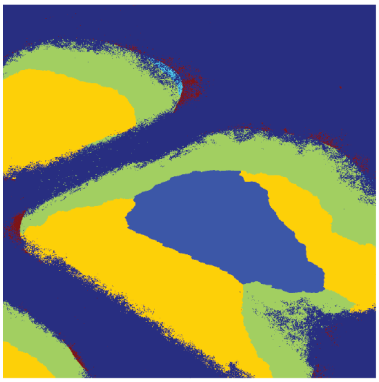
Perceptual attributes are used in [92] to classify macroscopic skin textures. Figure 4.4(a) shows exemplars of typical visual appearance in RCM skin images. Patches of size 150×150 are randomly sampled from the training images and used to create a labeled attribute dataset of 40000 training and 20000 test samples. Texton histograms of each attribute patch are used to train a neural network classifier. The trained attribute classifier performs with an accuracy of 97.2% on the test set. For test RCM images, a patch centered at each pixel of the RCM image can be labeled as one these perceptual attributes as shown in Figure 4.4(c). Integral histograms are used for computational speed of the input feature vectors corresponding to each pixel [149, 185]. We extend this method by using the histograms of



(a)



(b)



(c)

Figure 4.4: **Attributes-based approach.** (a) Perceptual Attributes. Exemplars of typical visual appearance (150 × 150 patches) in RCM skin images. (b) Input RCM skin image. (c) A patch centered at each pixel is labeled as one of the perceptual attributes.

the attributes labeled RCM image to train another neural network with an input layer, a hidden layers and an output layer with 6, 20 and 6 neurons, respectively.

4.2.4 Convolutional Neural Networks:

A convolutional neural network (CNN) typically consists of multiple convolutional, pooling and rectilinear linear units followed by a fully connected layer [100]. Our CNN architecture as shown in Figure 4.5 consists of following layers: The input RCM images are resized to 250 × 250 and given as input to the first convolutional layer which consists of 48 kernels of size 11 × 11 × 1 with a stride of 1 pixel. The filter output is followed by a max-pooling layer with stride of 4 pixels. The second convolutional layer takes the pooled output of first

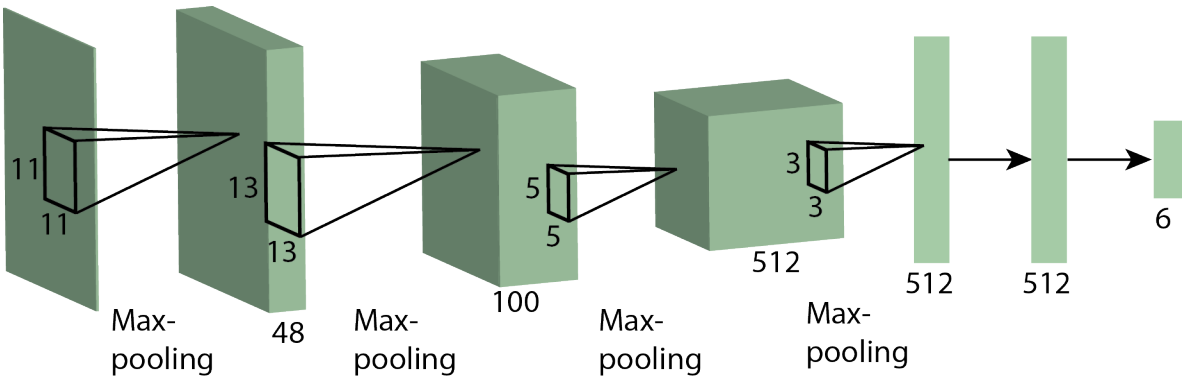


Figure 4.5: **Our CNN Architecture** consists of four convolutional layers with kernels of sizes 11, 13, 5 and 3. Each convolutional layer is followed by a max-pooling layer. The last layer of the network is a fully-connected layer with 512 neurons which gives output labels of the test image.

layer as input and filters it with 100 kernels of size 13×13. The third convolutional layers is connected to the pooled output of the previous layer with 512 kernels of size 5×5. The fourth convolutional layer has filter size 3×3. Finally, the fully-connected layers have 512 neurons each with 6 output neurons. We tried different variations of the CNN layers but they resulted in similar performance on the test data.

Since the RCM images dataset is limited, it results in over fitting of the CNN during the training phase and hence poor accuracy during the test phase. To overcome the problem of over-fitting, a network trained on a larger dataset such as Imagenet [157] can be use for transfer learning as opposed to training a new CNN from scratch. ImageNet is a dataset of natural images with 1000 classes and millions of images per class. It has been used to train several CNN architectures such as AlexNet [100], VGG-19 [171], GoogLeNet [179] and ResNet [75]. The networks learn rich feature representations from a variety of images. The authors of these architectures have made available the weights of the pre-trained CNNs. If a dataset has the same classes as ImageNet, then the network weights of the pre-trained CNNs on ImageNet can be used directly to predict the class of the images from the new dataset. If a dataset has very different classes from ImageNet, the pre-trained network weights can be used to extract the features vectors for each image. After extracting features from all the training images, a classifier like Support Vectors Machine or Deep Neural Network can be used to classify the images.

In our experiments, we use four different pre-trained networks which have been trained

on the ImageNet dataset:

1. **AlexNet** [100] reintroduced the CNNs in computer vision by winning the 2012 ILSVRC (ImageNet Large-Scale Visual Recognition Challenge) competition with an error rate of 15.4%. It consists of 5 convolutional and 3 fully connected layers. This architecture uses Rectifier Linear Unit (ReLU) layer for non-linearity functions instead of the traditional *tanh* layer used in neural networks. It also uses data augmentation to make the architecture robust to scale, size and orientation. Furthermore, dropout layer is used to avoid over-fitting.

2. **VGG-19** [171] reduced the error rate to 7.3% for the ILSVRC competition in 2014. It is a 19 layer deep convolution network which uses smaller filter sizes (3×3) for the convolution layers, hence increasing the depth and parameters of the network. It has been shown to perform better than AlexNet on ImageNet dataset.

3. **GoogLeNet** [179] is a deep convolution architecture which uses the concept of inception modules. It dropped the error rate to 6.7% for the ILSVRC competition in 2015. In AlexNet or VGG19, at each layer of the CNN we have to choose a filter size (for example 3×3, 5×5, 11×11 etc.). Its inception module allows convolution with multiple filters of different filter sizes at each layer. The filter outputs of multiple filter convolutions in each stage are concatenated. The network is designed such that the the number of parameters is small even though the complexity and computations have increased. Interestingly, the number of parameters is smaller than Alexnet by a factor of over 10x. The inception module performs better than a single convolution layer since it extracts multi-level features at the same time in each layer.

4. **ResNet** [75] (Deep Residual Network) is a 152 layer architecture which dropped the error rate to 3.6% for ILSVRC 2015. As the network depth is increased, gradient diminishes slightly during back propagation as it passes through each network layer. To overcome this problem of vanishing gradients, ResNet allows the gradient to pass backwards directly by skipping over all the intermediate layers such that it reaches the bottom layer without being diminished.

We use the pre-trained CNNs to extract feature vectors in two ways: globally and locally. In the global method, we resize each RCM image of size 1000×1000 to the input size of the pre-trained network (227×227 or 224×224). The last layer of the CNN is removed and an image can be passed through the rest of the network. For example, in VGG-19 model the last layer (1000-dimensional) can be removed and the fully connected layer (fc2) results in a 4096-dimesnional feature vector representation of an input image. Feature vectors of RCM images are used to train a deep neural network to classify the RCM images. For a test RCM image, the entire image is passed through the network which predicts the output class label. Since the image is resized, some texture information is lost which is reflected in the performance of the network.

To overcome the problem of information loss in resized images, we use a local patch-based method. During training phase, multiple random patches are extracted from the training images RCM image and each patch is assigned the same label as the image label. Patch size is same as that of the input layer of the pre-trained CNN. Feature vectors of these labeled patches are used to train a deep neural network. For a test RCM images, 50 random patches are extracted and each of it is passed through the network which predicts the output class label of each patch. The class label with maximum votes from all the 50 patch labels becomes the predicted test image label.

4.3 Experiments and Results

To evaluate the performance of algorithms for RCM skin classification, we perform k-fold cross-validation with $k = 5$ on our dataset. For each iteration, the training consists of 12 RCM stacks and the classifier is tested on three RCM stacks. Each RCM stack consists of 100 images. The ground truth for each image is obtained through manual labeling by a clinical expert. Each image is classified as one of the six skin layers (Figure 4.1). The following performance metrics are computed in each fold using the confusion matrix for the test sets: accuracy, sensitivity, specificity, precision and f-score [46]. Table 4.1 lists the average performance over five folds. Our proposed hybrid deep learning method outperforms other methods. Using the hybrid deep learning approach in Section 4.2.2, average accuracy on test sets is 0.8173 (or 81.73%). If we replace the multi-layer neural network by SVM classifier, the accuracy reduces to 0.75.

The perceptual attribute based approach described in Section 4.2.3 gives pixel level attribute labels. An example of a test image labeled with perceptual attributes is shown in Figure 4.4. The histogram of attribute labels given as a feature to a multi-layer neural network results in an accuracy of 0.7120. The CNN architecture proposed in Section 4.2.4 performs poorly with average test accuracy of 0.5147. We designed the architecture empirically and tried variations of different layers and kernel sizes. The network was implemented in MATLAB using MatConvNet library [135, 186]. The CNNs learn the kernel weights automatically from the labeled training data. Since our datasets is moderate size, it may be difficult for the network to converge and learn the correct filter weights. Further, the higher layers of CNN also learn complex features apart from the texture such as external contour shape, which is not useful for skin texture classification of RCM images.

We also compare our method to the unsupervised texton based approach in [174] which gives an accuracy of 0.5395%. The patch based bag-of-features approach proposed in [72] followed by a SVM classifier and deep neural network give an accuracy of 0.5413 and 0.7993, respectively. In addition to accuracy, the hybrid deep learning method also results in the best sensitivity, specificity, precision and f-score among all the methods.

	Accuracy ± standard deviation	**Sensitivity**	**Specificity**	**Precision**	**F-score**
Unsupervised texton-based approach [18]	0.5395 ± 0.03	0.5270	0.90	0.5422	0.4883
Patch based bag-of-features approach followed by SVM classifier [20]	0.5413 ± 0.25	0.3811	0.9020	0.5128	0.3721
Patch based bag-of-features approach followed by a Neural Network classifier	0.7993 ± 0.05	0.6918	0.9587	0.6990	0.6842
Convolutional Neural Network	0.5134 ± 0.07	0.20	0.8392	0.60	0.1653
Attribute approach (extension of [23])	0.7120 ± 0.04	0.5564	0.9392	0.5711	0.5520
Texton approach followed by SVM classifier	0.75 ± 0.03	0.6102	0.9482	0.6346	0.6013
Hybrid deep learning	**0.8173 ± 0.05**	**0.7174**	**0.9620**	**0.7236**	**0.7104**

Table 4.1: **Comparison of different approaches for RCM skin image classification.** Our proposed hybrid deep learning outperforms other methods.

	Accuracy	
	Global Method	Local Patch-based Method
Hybrid Deep Learning	79.47	81.67
CNN: AlexNet	80.87	84.60
CNN: VGG-19	82.60	85.53
CNN: GoogLeNet	66.00	83.80
CNN: ResNet	69.67	85.87

Table 4.2: **Comparison of CNN approach for RCM Image Classification.** Using pre-trained CNNs on the entire image in the global method either marginally improves or reduces the accuracy. Using patch-based method, improves the overall accuracy.

	OE	130	25	0	0	0	1
	SC	14	67	17	5	0	0
Output	SG	0	22	53	18	0	0
Class	SS	2	1	20	146	37	0
	SB	0	0	2	32	89	42
	PD	6	0	0	1	33	737
		OE	SC	SG	SS	SB	PD
		Input Class					

Table 4.3: **Confusion Matrix.** Note that the mislabeling is between adjoining skin layers. Also see examples in Figure 4.7.

Even though training a CNN from scratch using RCM images results in a poor classification, using the pre-trained CNNs can improves the classification accuracy. In the global method, where the entire RCM image is resized and used as CNN input, the accuracy either marginally improves or reduces. As shown in Table 4.2, using AlexNet or VGG-19, marginally improves the accuracy in comparison to the hybrid deep learning approach (80.87% and 82% as compared to 79.47%). On the other hand, much deeper networks: GoogleNet and ResNet have reduced accuracy (66% and 69.97%). In the local patch-based method, training is done on random patches from training images and the label of the test image is assigned by a maximum vote of 50 random patches. Using this method, we observe that using pre-trained CNNs results in better acuuracy as compared to the hybrid deep learning approach. ResNet gives the best accuracy (85.87%), followed by VGG-19 (85.53%).

In most stacks, the mislabeling is observed between adjoining layers as the images transition from one skin layer to another. Table 4.3 shows the confusion matrix for all the test

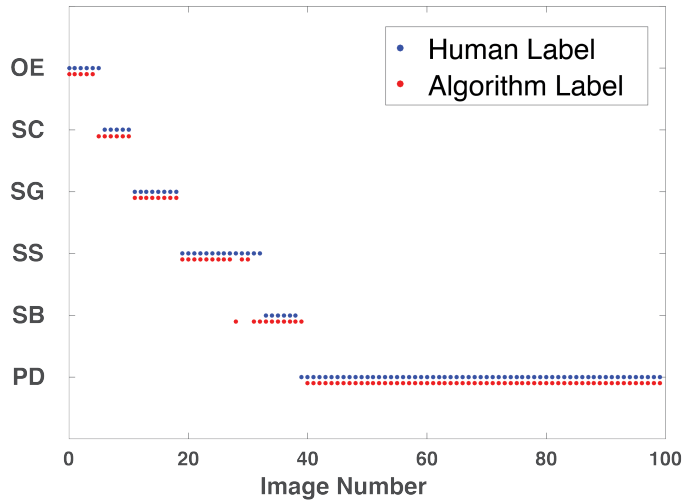


Figure 4.6: **Example of RCM test stack labeling.** This test stack has 100 RCM images. Each image, represented by a dot, is labeled as one of the skin layers by human (blue dots) and our algorithm (red dots). Note that the mislabeling is between adjoining skin layers. Also see examples in Figure 4.7.

	Accuracy	
	Global Method	Local Patch-based Method
Hybrid Deep Learning	83.40	85.93
CNN: AlexNet	84.87	88.87
CNN: VGG-19	86.47	89.20
CNN: GoogLeNet	68.87	88.00
CNN: ResNet	72.07	89.87

Table 4.4: **Classification accuracy by ignoring mislabeling at the transition regions.**

stacks. Figure 4.7 shows examples of mislabeled images in the first column. The correctly labeled images are shown in the second and third columns. Figure 4.7(a) was labeled as SC by the algorithm but as OE by the human expert. Figures 4.7(b) and (c) show examples correctly labeled by the algorithm as SC and OE, respectively. Such transitions may be ambiguous to a clinical expert as well. Table 4.4 shows the classification accuracy by allowing mislabeling at the ambiguous transition regions. Patch-based CNN approach using ResNet gives the best accuracy (89.87%), followed by VGG-19 (89.20%).

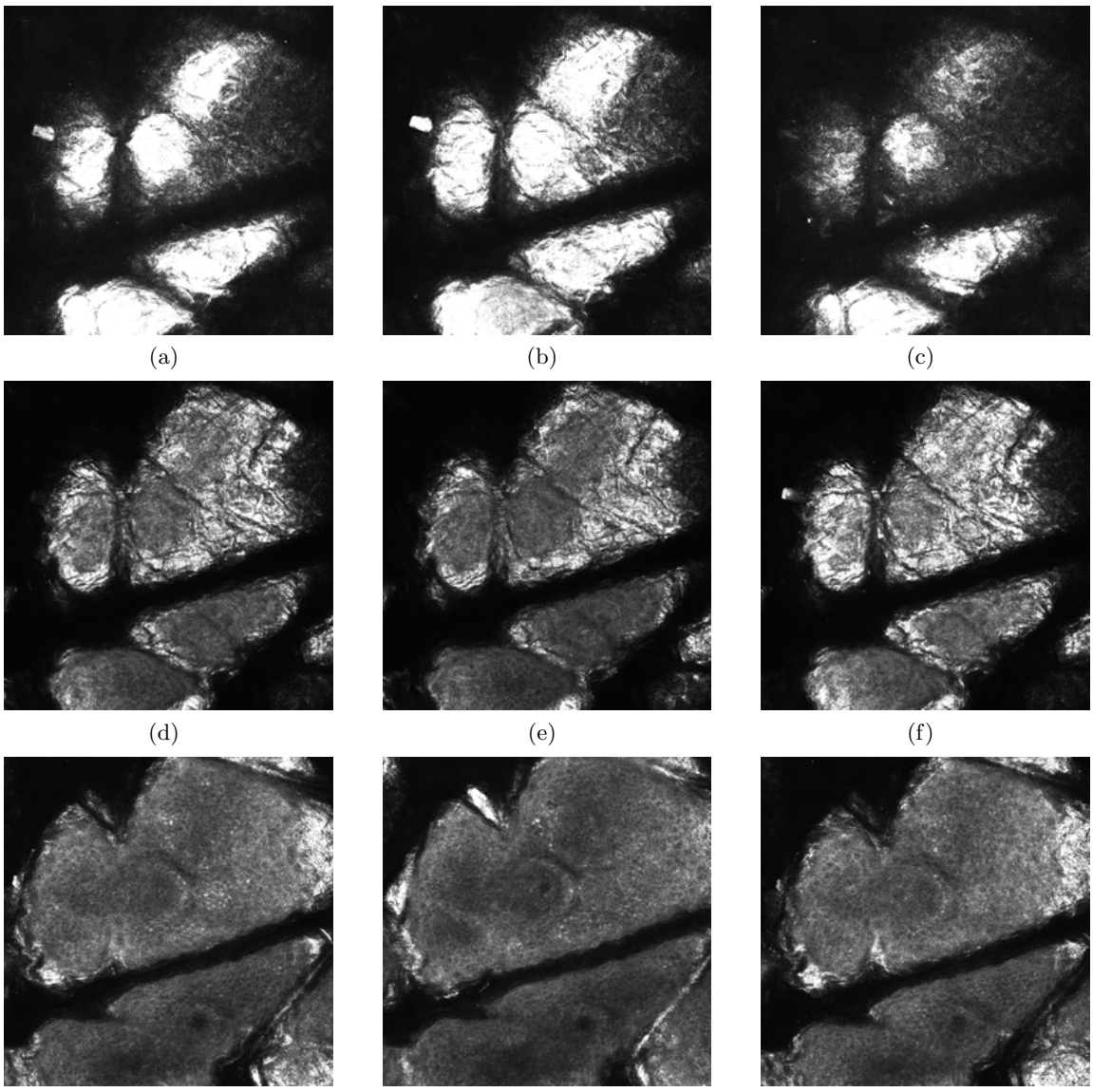


Figure 4.7: **Examples of mislabeled RCM images.** Ambiguity of appearance in transition regions between skin layers. *First Column:* Examples of mislabeled images. (a) Human label:OE, Automated label:SC. (d) Human label:SC, Automated label:SS. (g) Human label:SS, Automated label:SB. *Second Column:* Correctly labeled: (b) SC (e) SS (h) SB. *Third Column:* Correctly labeled: (c) OE (f) SC (i) SS.

4.4 Conclusions

In this chapter we introduce a hybrid deep learning approach for automatically labeling RCM skin images. Texton-based features are obtained using a fixed multi-resolution, multi-orientation filter bank to train a deep neural network. We compare our method with a suite of texture recognition methods and show that it outperforms the state-of-the-art with a test accuracy of 81.73%. We demonstrate that smaller training datasets are insufficient for CNN training and feature extraction is essential in such cases. Using patch-based approach and pre-trained CNNs for feature extraction, we achieve the classification accuracy of 85.53%. Furthermore, we highlight the ambiguity in labeling at transition regions and achieve 89.87% accuracy by allowing mislabeling at the transition regions.

Chapter 5

Conclusions and Future Work

In this dissertation, we propose deep learning based approaches to address different problems in the field of dermatology. In Chapter 2, we presented an approach using convolutional neural network for texture transfer between facial images (FcaeTex) to visualize the effects of aging, dun exposure or skin treatments. We introduce regularizations and network losses that can preserve the identity of a person while transferring texture from another person's face. Besides the qualitative visual evaluation, we also proposed two measures of quantitative evaluation: landmark error and texture similarity. We compare our results with state-of-the-art methods and demonstrate better identity-preserving texture transfer.

In Chapter 3, we propose methods to link skin microbiome and skin appearance. Measuring skin microbiome through sequencing is time consuming whereas skin appearance can be instantaneously captured. We present a computational appearance model for images in ultraviolet and fluorescence modalities. The intersection of appearance and microbiome clusters reveals three interesting cases where microbiome clusters can be predicted from the appearance clusters with high probabilities. We also propose appearance-driven multiview co-clustering to discover associations of microbiome and appearance. In these experiments, our dataset is limited to 48 subjects. By increasing the sample size in the dataset and applying advanced techniques, this work has potential to make more scientific discoveries.

In Chapter 4, we propose a hybrid deep learning approach as well as a patch-based CNN approach to classify microscopic skin images and measure thickness of skin layers. We achieve classification accuracy of 85.53% and highlight the errors at transition regions. By ignoring the transition regions, we achieve 89.87% classification accuracy. In this work, we have image labels only from a single expert. Also, our dataset had 1500 RCM images. Collecting more data and getting it labeled from more than one expert can lead to further

improvements in classification accuracy and better insights about labeling variability.

Deep learning methods proposed in this research can be extended to several other dermatology applications. In Chapter 4, we used hybrid deep learning and CNNs for classification of microscopic skin images. CNNs have also been used for classification of skin cancers from macroscopic images [45]. In future work, these techniques can also be extended for fine-grained classification of skin diseases, such as acne severity. Furthermore, CNNs can also localize skin features using the labeled data [41, 144]. This will essentially enable the network architecture to computationally "see" what dermatologists see and get insights on why an image belongs to a certain category. An interesting application of this work can be to train the dermatologists, who are just starting their career.

Artificial intelligence (AI) has led to a boom in virtual assistants and recommender systems in the last decade. These AI tools recommend new products to a user based on his/her past product history and choices made by similar users. Similar skin assistants have been designed to analyze skin and recommend essential products [1] for maintaining a healthy skin and prevent it from damage due to environmental factors. In future, the existing skin product recommender systems can be improved and quantitatively evaluated. Deep learning methods have also been used for super-resolution methods, where a high-resolution image is obtained from a low-resolution image [37, 109]. These methods can be used to obtaining high-resolution images for skin analysis. Since skin datasets are unique and many times limited in quantity, it is also worth exploring generative adversarial networks to generate artificial skin datasets.

Deep learning has led to remarkable progress in several artificial intelligence applications. Quantitative dermatology can leverage from these methods to assist dermatologists and gain a better understanding of skin.

Appendix A

LM Filter Bank

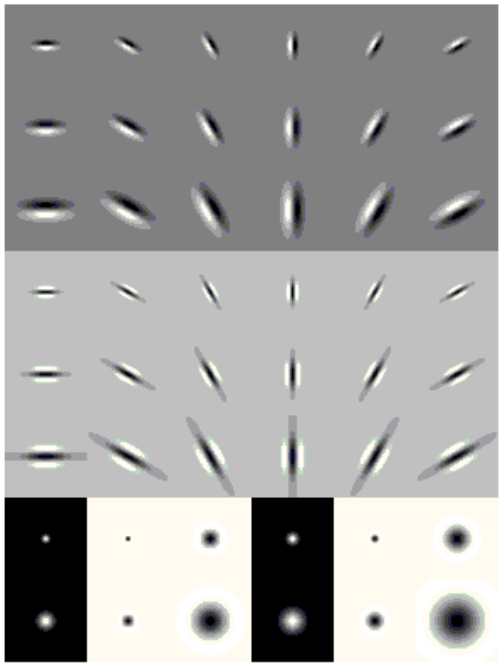


Figure A.1: **LM Filter Bank** [113] is comprised of 48 filters. These filters include 36 first and second order derivative of Gaussain filters (6 orientations, 3 scales each), 8 Laplacian of Gaussain filters and 4 Gaussian filters.

Appendix B

AMCO: FLUO and UV datasets

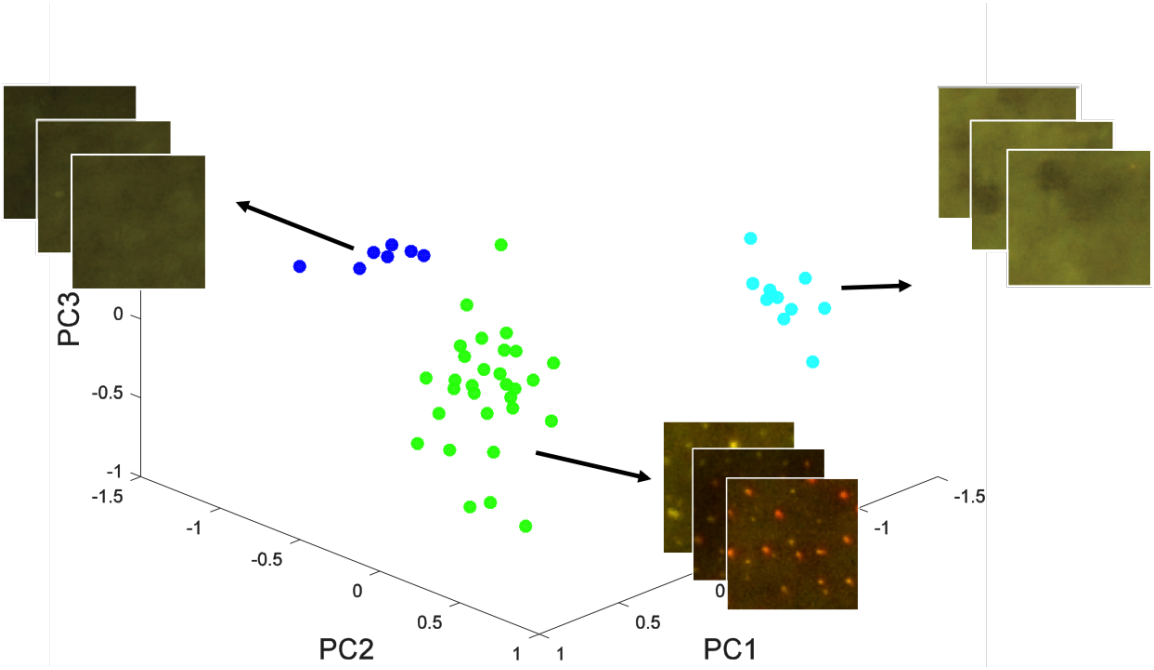


Figure B.1: AMCO: appearance(FLUO)-microbiome. The data in secondary (microbiome) space (projected to 3D for visualization). The appearance clusters in FLUO modality group together in secondary space after AMCO (each color represents an appearance cluster).

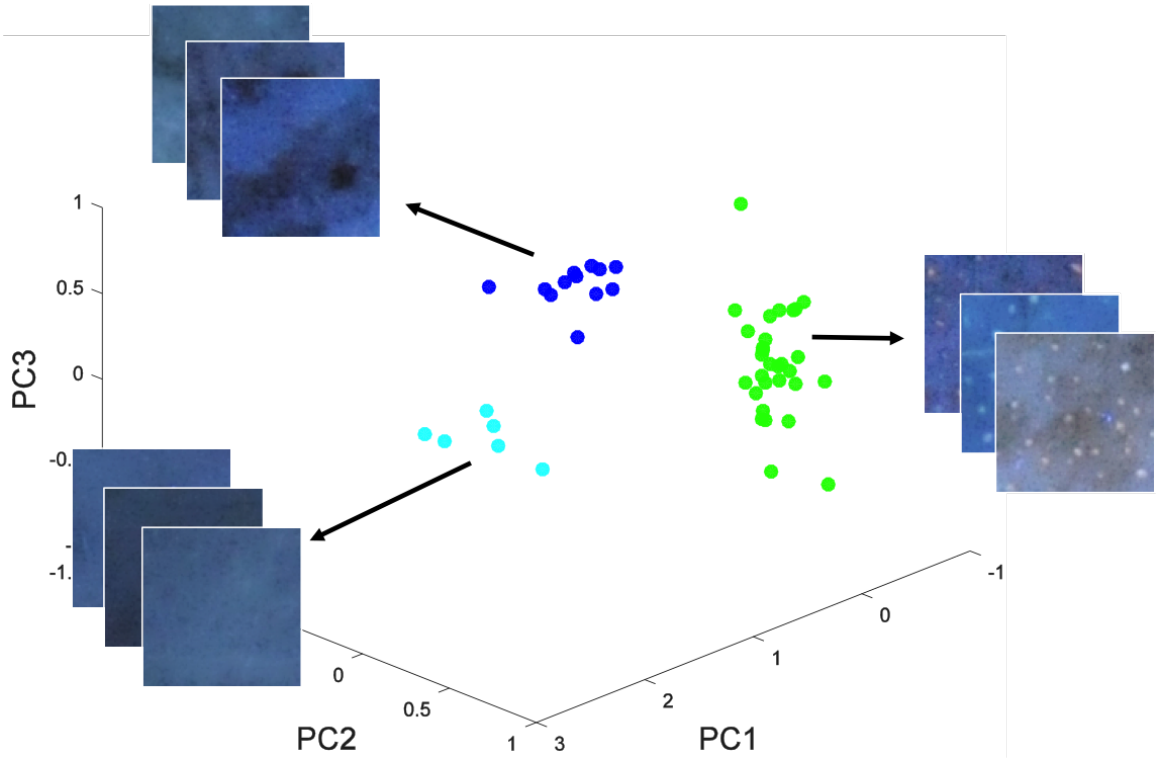


Figure B.2: AMCO: appearance(UV)-microbiome. The data in secondary (microbiome) space (projected to 3D for visualization). The appearance clusters in UV modality group together in secondary space after AMCO (each color represents an appearance cluster).

Appendix C

AMCO: animals with attributes dataset

Figure C.1: Animal with attributes dataset: before AMCO. The data in secondary (attribute) space (projected to 3D for visualization). The appearance clusters do not group together in secondary space (each color represents an appearance cluster).

Figure C.2: Animal with attributes dataset: after AMCO. The data in secondary (attribute) space (projected to 3D for visualization). The appearance clusters are visually distinguishable in secondary space after AMCO (each color represents an appearance cluster).

References

[1] Play skin analyzer. https://skinadvisor.olay.co.uk/en-GB/. Accessed: 2017-08-30.

[2] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid. Label-embedding for attribute-based classification. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 819–826, June 2013.

[3] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *arXiv preprint arXiv:1511.00561*, 2015.

[4] John S Barbieri, Caroline A Nelson, William D James, David J Margolis, Ryan Littman-Quinn, Carrie L Kovarik, and Misha Rosenbach. The reliability of teledermatology to triage inpatient dermatology consultations. *JAMA dermatology*, 150(4):419–424, 2014.

[5] Thaddeus Beier and Shawn Neely. Feature-based image metamorphosis. In *ACM SIGGRAPH Computer Graphics*, volume 26, pages 35–42. ACM, 1992.

[6] ET Bell. The relation of portal cirrhosis to hemochromatosis and to diabetes mellitus. *Diabetes*, 4(6):435–446, 1955.

[7] Steffen Bickel and Tobias Scheffer. Multi-view clustering. In *ICDM*, volume 4, pages 19–26, 2004.

[8] Xiao Cai, Feiping Nie, and Heng Huang. Multi-view k-means clustering on big data. In *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*, pages 2598–2604. AAAI Press, 2013.

[9] Xiao Cai, Feiping Nie, Heng Huang, and F. Kamangar. Heterogeneous image feature integration via multi-modal spectral clustering. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1977–1984, June 2011.

[10] Jeffrey P Callen, Joseph Jorizzo, Kenneth E Greer, Neal Penneys, Warren W Piette, and John J Zone. *Dermatological signs of internal disease*. WB Saunders New York, 1995.

[11] Piergiacomo Calzavara-Pinton, Caterina Longo, Marina Venturini, Raffaella Sala, and Giovanni Pellacani. Reflectance confocal microscopy for in vivo skin imaging. *Photochemistry and photobiology*, 84(6):1421–1430, 2008.

[12] Xiaochun Cao, Changqing Zhang, Huazhu Fu, Si Liu, and Hua Zhang. Diversity-induced multi-view subspace clustering. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[13] Kimberly A Capone, Scot E Dowd, Georgios N Stamatas, and Janeta Nikolovski. Diversity of the human skin microbiome early in life. *Journal of Investigative Dermatology*, 131(10):2026 – 2032, 2011.

[14] Kamalika Chaudhuri, Sham M Kakade, Karen Livescu, and Karthik Sridharan. Multi-view clustering via canonical correlation analysis. In *Proceedings of the 26th annual international conference on machine learning*, pages 129–136. ACM, 2009.

[15] Shizhi Chen, YingLi Tian, Qingshan Liu, and Dimitris N. Metaxas. Recognizing expressions from face and body gesture by temporal normalized motion and appearance features. *Image and Vision Computing*, 31(2):175 – 185, 2013. Affect Analysis In Continuous Input.

[16] YE Chen and H Tsao. The skin microbiome: Current perspectives and future challenges. *Journal of the American Academy of Dermitology*, 69(1):143 – 155, 2013.

[17] Mircea Cimpoi, Subhransu Maji, and Andrea Vedaldi. Deep filter banks for texture recognition and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3828–3836, 2015.

[18] Brittany G Craiglow, Jack S Resneck, Anne W Lucky, Robert Sidbury, Albert C Yan, Steven D Resnick, and Richard J Antaya. Pediatric dermatology workforce shortage: perspectives from academia. *Journal of the American Academy of Dermatology*, 59(6):986–989, 2008.

[19] G. O. Cula, P. R. Bargo, A. Nkengne, and N. Kollias. Assessing facial wrinkles: automatic detection and quantification. *Skin Research & Technology*, 19(1):e243 – e251, 2013.

[20] G. Oana Cula, J. Kristin Dana, P. Frank Murphy, and K. Babar Rao. Skin texture modeling. *International Journal of Computer Vision*, 62(1):97–119, 2005.

[21] Gabriela O. Cula, Paulo R. Bargo, and Nikiforos Kollias. Imaging inflammatory acne: lesion detection and tracking. *Proc. SPIE*, 7548:75480I–75480I–7, 2010.

[22] O. G. Cula and K. J. Dana. Compact representation of bidirectional texture functions. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1:1041–1067, December 2001.

[23] O. G. Cula and K. J. Dana. Recognition methods for 3d textured surfaces. *Proceedings of SPIE Conference on Human Vision and Electronic Imaging VI*, 4299:209–220, January 2001.

[24] O. G. Cula, K. J. Dana, F. P. Murphy, and B. K. Rao. Skin texture modeling. *International Journal of Computer Vision*, 62(1/2):97–119, April/May 2005.

[25] O. G. Cula, K. J. Dana, D. K. Pai, and D. Wang. Polarization multiplexing and demultiplexing for appearance-based modeling.

[26] O. G. Cula, K. J. Dana, D. K. Pai, and D. Wang. Polarization multiplexing for bidirectional imaging. *IEEE Conference on Computer Vision and Pattern Recognition CVPR*, pages 1116–1123, June 2005.

[27] Oana G. Cula and Kristin J. Dana. 3d texture recognition using bidirectional feature histograms. *International Journal of Computer Vision*, 59:2004, 2004.

[28] Oana Gabriela Cula and Kristin J. Dana. Texture for appearance models in computer vision and graphics. In Majid Mirmehdi, Xianghua Xie, and Jasjit Suri, editors, *Handbook of Texture Analysis*. Imperial College Press, 2008.

[29] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE, 2005.

[30] Kristin J. Dana, Oana G. Cula, and Jing Wang. Surface detail in computer models. *Image and Vision Computing*, 25(7):1037 – 1049, 2007.

[31] Jeremy S De Bonet. Multiresolution sampling procedure for analysis and synthesis of texture images. In *Proceedings of the 24th annual conference on Computer graphics and interactive techniques*, pages 361–368. ACM Press/Addison-Wesley Publishing Co., 1997.

[32] Virginia R de Sa. Spectral clustering with two views. In *ICML workshop on learning with multiple views*, pages 20–27, 2005.

[33] Virginia R De Sa, Patrick W Gallagher, Joshua M Lewis, and Vicente L Malave. Multi-view kernel construction. *Machine learning*, 79(1-2):47–71, 2010.

[34] Les Dethlefsen, Sue Huse, Mitchell L Sogin, and David A Relman. The pervasive effects of an antibiotic on the human gut microbiota, as revealed by deep 16s rrna sequencing. *PLoS biology*, 6(11):e280, 2008.

[35] Inderjit S Dhillon. Co-clustering documents and words using bipartite spectral graph partitioning. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 269–274. ACM, 2001.

[36] TL Diepgen and V Mahler. The epidemiology of skin cancer. *British Journal of Dermatology*, 146(s61):1–6, 2002.

[37] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *European Conference on Computer Vision*, pages 184–199. Springer, 2014.

[38] Craig Donner, Tim Weyrich, Eugene d'Eon, Ravi Ramamoorthi, and Szymon Rusinkiewicz. A layered, heterogeneous reflectance model for acquiring and rendering human skin. *ACM Trans. Graph.*, 27(5):140:1–140:12, December 2008.

[39] Alexey Dosovitskiy, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. Discriminative unsupervised feature learning with convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 766–774, 2014.

[40] Vincent Dumoulin, Jonathon Shlens, and Manjunath Kudlur. A learned representation for artistic style. *CoRR*, abs/1610.07629, 2016.

[41] Thibaut Durand, Nicolas Thome, and Matthieu Cord. Weldon: Weakly supervised learning of deep convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4743–4752, 2016.

[42] Alexei A Efros and William T Freeman. Image quilting for texture synthesis and transfer. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 341–346. ACM, 2001.

[43] Alexei A Efros and Thomas K Leung. Texture synthesis by non-parametric sampling. In *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, volume 2, pages 1033–1038. IEEE, 1999.

[44] Logan Engstrom. Fast style transfer. https://github.com/lengstrom/fast-style-transfer/, 2016. commit 1424228.

[45] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118, 2017.

[46] Tom Fawcett. An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874, 2006.

[47] L. Fei-Fei and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 2, pages 524–531 vol. 2, June 2005.

[48] Sorel Fitz-Gibbon, Shuta Tomida, Bor-Han Chiu, Lin Nguyen, Christine Du, Minghsun Liu, David Elashoff, Marie C Erfe, Anya Loncaric, Jenny Kim, Robert L Modlin, Jeff F Miller, Erica Sodergren, Noah Craft, George M Weinstock, and Huiying Li. Propionibacterium acnes strain populations in the human skin microbiome associated with acne. *Journal of Investigative Dermatology*, 133(9):2152 – 2160, 2013.

[49] Vincent Foulongne, Virginie Sauvage, Charles Hebert, Olivier Dereure, Justine Cheval, Meriadeg Ar Gouilh, Kevin Pariente, Michel Segondy, Ana Burguire, Jean-Claude Manuguerra, Valrie Caro, and Marc Eloit. Human skin microbiota: High diversity of dna viruses identified on the human skin by high throughput sequencing. *PLoS ONE*, 7(6):1 – 11, 2012.

[50] M Garcia-Garcera, K Garcia-Etxebarria, M Coscolla, A Latorre, and F Calafell. A new method for extracting skin microbes allows metagenomic analysis of whole-deep skin. *PLOS ONE*, 8(9), 2013.

[51] Dan Gareau, Ricky Hennessy, Eric Wan, Giovanni Pellacani, and Steven L Jacques. Automated detection of malignant features in confocal microscopy on superficial spreading melanoma versus nevi. *Journal of biomedical optics*, 15(6):061713–061713, 2010.

[52] Leon A Gatys, Matthias Bethge, Aaron Hertzmann, and Eli Shechtman. Preserving color in neural artistic style transfer. *arXiv preprint arXiv:1606.05897*, 2016.

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

[54] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Texture synthesis using convolutional neural networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*, NIPS'15, pages 262–270, Cambridge, MA, USA, 2015. MIT Press.

[55] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[56] Leon A Gatys, Alexander S Ecker, Matthias Bethge, Aaron Hertzmann, and Eli Shechtman. Controlling perceptual factors in neural style transfer. *arXiv preprint arXiv:1611.07865*, 2016.

[57] T. George and S. Merugu. A scalable collaborative filtering framework based on co-clustering. In *Data Mining, Fifth IEEE International Conference on*, pages 4 pp.–, Nov 2005.

[58] Abhijeet Ghosh, Tim Hawkins, Pieter Peers, Sune Frederiksen, and Paul Debevec. Practical modeling and acquisition of layered facial reflectance. *ACM Trans. Graph.*, 27(5):139:1–139:10, December 2008.

[59] Catherine S Giess, Sughra Raza, and Robyn L Birdwell. Distinguishing breast skin lesions from superficial breast parenchymal lesions: diagnostic criteria, imaging characteristics, and pitfalls. *Radiographics*, 31(7):1959–1972, 2011.

[60] Amir Globerson and Sam T Roweis. Metric learning by collapsing classes. In *Advances in neural information processing systems*, pages 451–458, 2005.

[61] Hugh Morris Gloster Jr, Lauren E Gebauer, and Rachel L Mistur. Cutaneous manifestations of addisons disease. In *Absolute Dermatology Review*, pages 169–169. Springer, 2016.

[62] Jacob Goldberger, Geoffrey E Hinton, Sam T Roweis, and Ruslan Salakhutdinov. Neighbourhood components analysis. In *Advances in neural information processing systems*, pages 513–520, 2004.

[63] Yunchao Gong, Qifa Ke, Michael Isard, and Svetlana Lazebnik. A multi-view embedding space for modeling internet images, tags, and their semantics. *International journal of computer vision*, 106(2):210–233, 2014.

[64] Michael Goodfellow and Erko Stackebrandt. *Nucleic acid techniques in bacterial systematics*. J. Wiley, 1991.

[65] Elizabeth A Grice. The skin microbiome: potential for novel diagnostic and therapeutic approaches to cutaneous disease. In *Seminars in cutaneous medicine and surgery*, volume 33, pages 98–103. Frontline Medical Communications, 2014.

[66] Elizabeth A Grice, Heidi H Kong, Sean Conlan, Clayton B Deming, Joie Davis, Alice C Young, Gerard G Bouffard, Robert W Blakesley, Patrick R Murray, Eric D Green, et al. Topographical and temporal diversity of the human skin microbiome. *science*, 324(5931):1190–1192, 2009.

[67] Elizabeth A Grice, Heidi H Kong, Gabriel Renaud, Alice C Young, Gerard G Bouffard, Robert W Blakesley, Tyra G Wolfsberg, Maria L Turner, and Julia A Segre. A diversity profile of the human skin microbiota. *Genome research*, 18(7):1043–1050, 2008.

[68] Elizabeth A. Grice and Julia A. Segre. The skin microbiome. *Nature Reviews Microbiology*, 9(4):244 – 253, 2011.

[69] The NIH HMP Working Group, Jane Peterson, Susan Garges, Maria Giovanni, Pamela McInnes, Lu Wang, Jeffery A. Schloss, Vivien Bonazzi, Jean E. McEwen, Kris A. Wetterstrand, Carolyn Deal, Carl C. Baker, Valentina Di Francesco, T. Kevin Howcroft, Robert W. Karp, R. Dwayne Lunsford, Christopher R. Wellington, Tsegahiwot Belachew, Michael Wright, Christina Giblin, Hagit David, Melody Mills, Rachelle Salomon, Christopher Mullins, Beena Akolkar, Lisa Begg, Cindy Davis, Lindsey Grandison, Michael Humble, Jag Khalsa, A. Roger Little, Hannah Peavy, Carol Pontzer, Matthew Portnoy, Michael H. Sayre, Pamela Starke-Reed, Samir Zakhari, Jennifer Read, Bracie Watson, and Mark Guyer. The NIH Human Microbiome Project. *Genome Research*, 19(12):2317–2323, December 2009.

[70] Dongyan Guo, Jian Zhang, Xinwang Liu, Ying Cui, and Chunxia Zhao. Multiple kernel learning based multi-view spectral clustering. In *Pattern Recognition (ICPR), 2014 22nd International Conference on*, pages 3774–3779, Aug 2014.

[71] Samuel Hames, Marco Ardigò, H Peter Soyer, Andrew P Bradley, and Tarl W Prow. Automated segmentation of skin strata in reflectance confocal microscopy depth stacks. *bioRxiv*, page 022137, 2015.

[72] Samuel C Hames, Marco Ardigo, H Peter Soyer, Andrew P Bradley, and Tarl W Prow. Anatomical skin segmentation in reflectance confocal microscopy with weak labels. In *Digital Image Computing: Techniques and Applications (DICTA), 2015 International Conference on*, pages 1–8. IEEE, 2015.

[73] Byungkwan Han, Byungjo Jung, J. Stuart Nelson, and Eung-Ho Choi. Analysis of facial sebum distribution using a digital fluorescent imaging system. *Journal of Biomedical Optics*, 12(1):014006–014006–6, 2007.

[74] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

[75] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[76] David J Heeger and James R Bergen. Pyramid-based texture analysis/synthesis. In *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*, pages 229–238. ACM, 1995.

[77] Viday A Heffner, Valerie B Lyon, David C Brousseau, Kristin E Holland, and Kenneth Yen. Store-and-forward teledermatology versus in-person visits: a comparison in pediatric teledermatology clinic. *Journal of the American Academy of Dermatology*, 60(6):956–961, 2009.

[78] Rainer Hofmann-Wellenhof, Giovanni Pellacani, Joseph Malvehy, and H Peter Soyer. *Reflectance confocal microscopy for skin diseases*. Springer Science & Business Media, 2012.

[79] Steven L. Jacques, Jessica C. Ramella-Roman, and Ken Lee. Imaging skin pathology with polarized light. *Journal of Biomedical Optics*, 7(3):329–340, 2002.

[80] William D James, Dirk Elston, and Timothy Berger. *Andrew's Diseases of the Skin E-Book: Clinical Dermatology*. Elsevier Health Sciences, 2011.

[81] Sanford James A. and Gallo Richard L. Review: Functions of the skin microbiota in health and disease. *Seminars in Immunology*, 25(Microbiota and the immune system, an amazing mutualism forged by co-evolution):370 – 377, 2013.

[82] Justin Johnson. neural-style. https://github.com/jcjohnson/neural-style, 2015.

[83] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, 2016.

[84] S. Jones and Ling Shao. Unsupervised spectral dual assignment clustering of human actions in context. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 604–611, June 2014.

[85] Bela Julesz. Textons, the elements of texture perception, and their interactions. *Nature*, 290(5802):91–97, 1981.

[86] Byungjo Jung, Bernard Choi, Anthony J. Durkin, Kristen M. Kelly, and J. Stuart Nelson. Characterization of port wine stain skin erythema and melanin content using cross-polarized diffuse reflectance imaging. *Lasers in Surgery and Medicine*, 34(2):174–181, 2004.

[87] Chante Karimkhani, Robert P Dellavalle, Luc E Coffeng, Carsten Flohr, Roderick J Hay, Sinéad M Langan, Elaine O Nsoesie, Alize J Ferrari, Holly E Erskine, Jonathan I Silverberg, et al. Global skin disease morbidity and mortality: an update from the global burden of disease study 2013. *JAMA dermatology*, 153(5):406–412, 2017.

[88] Andrej Karpathy, George Toderici, Sachin Shetty, Tommy Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 1725–1732. IEEE, 2014.

[89] H. Kato and T. Harada. Image reconstruction from bag-of-visual-words. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 955–962, June 2014.

[90] Hang Kaur, Parneet abd Zhang and Kristin J Dana. Photo-realistic facial texture transfer.

[91] P. Kaur, K. J. Dana, G. O. Cula, and C. Mack. Hybrid deep learning for reflectance confocal microscopy skin images. In *2016 23rd International Conference on Pattern Recognition*, Dec 2016.

[92] Parneet Kaur, Kristin Dana, and Gabriela Cula. From photography to microbiology: Eigenbiome models for skin appearance. In *BioImage Computing Workshop, Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*. IEEE, 2015.

[93] Alexa Boer Kimball and Jack S Resneck. The us dermatology workforce: a specialty remains in shortage. *Journal of the American Academy of Dermatology*, 59(5):741–745, 2008.

[94] S Koller, M Wiltgen, V Ahlgrimm-Siess, W Weger, R Hofmann-Wellenhof, E Richtig, J Smolle, and A Gerger. In vivo reflectance confocal microscopy: automated diagnostic image analysis of melanocytic skin tumours. *Journal of the European Academy of Dermatology and Venereology*, 25(5):554–558, 2011.

[95] N. Kollias. *Bioengineering of the Skin: Skin Imaging and Analysis*. CRC Press, second edition, 2004.

[96] Heidi H Kong and Julia A Segre. Skin microbiome: Looking back to move forward. *Journal of Investigative Dermatology*, 132(3 part 2):933 – 939, 2012.

[97] Konstantin Korotkov and Rafael Garcia. Computerized analysis of pigmented skin lesions: A review. *Artificial Intelligence in Medicine*, 56(2):69 – 90, 2012.

[98] Kivanc Kose, Christi Alessi-Fox, Melissa Gill, Jennifer G Dy, Dana H Brooks, and Milind Rajadhyaksha. A machine learning method for identifying morphological patterns in reflectance confocal microscopy mosaics of melanocytic skin lesions in-vivo. In *SPIE BiOS*, pages 968908–968908. International Society for Optics and Photonics, 2016.

[99] Aravind Krishnaswamy and Gladimir VG Baranoski. A biophysically-based spectral model of light interaction with human skin. In *Computer Graphics Forum*, volume 23, pages 331–340. Wiley Online Library, 2004.

[100] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[101] Abhishek Kumar and Hal Daumé. A co-training approach for multi-view spectral clustering. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 393–400, 2011.

[102] Abhishek Kumar, Piyush Rai, and Hal Daume. Co-regularized multi-view spectral clustering. In *Advances in Neural Information Processing Systems*, pages 1413–1421, 2011.

[103] Sila Kurugol, Kivanc Kose, Brian Park, Jennifer G Dy, Dana H Brooks, and Milind Rajadhyaksha. Automated delineation of dermal–epidermal junction in reflectance confocal microscopy image stacks of human skin. *Journal of Investigative Dermatology*, 135(3):710–717, 2015.

[104] Vivek Kwatra, Arno Schödl, Irfan Essa, Greg Turk, and Aaron Bobick. Graphcut textures: image and video synthesis using graph cuts. In *ACM Transactions on Graphics (ToG)*, volume 22, pages 277–286. ACM, 2003.

[105] M Lai, I Oru, and JJ Barton. The role of skin texture and facial shape in representations of age and identity. *Cortex: A Journal Devoted to the Study of the Nervous System & Behavior*, 49(1):252 – 265, 2013.

[106] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Attribute-based classification for zero-shot visual object categorization. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 36(3):453–465, 2014.

[107] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Computer vision and pattern recognition, 2006 IEEE computer society conference on*, volume 2, pages 2169–2178. IEEE, 2006.

[108] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[109] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. *arXiv preprint arXiv:1609.04802*, 2016.

[110] Daniel D. Lee and H. Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, October 1999.

[111] Daniel D. Lee and H. Sebastian Seung. Algorithms for non-negative matrix factorization. In *In NIPS*, pages 556–562. MIT Press, 2000.

[112] Mihee Lee, Haipeng Shen, Jianhua Z Huang, and JS Marron. Biclustering via sparse singular value decomposition. *Biometrics*, 66(4):1087–1095, 2010.

[113] T. Leung and J. Malik. Representing and recognizing the visual appearance of materials using three-dimensional textons. *International Journal of Computer Vision*, 43(1):29–44, 2001.

[114] W Levinson. *Normal Flora of the Skin.* McGraw-Hill, 2012.

[115] Chuan Li and Michael Wand. Precomputed real-time texture synthesis with markovian generative adversarial networks. In *European Conference on Computer Vision*, pages 702–716. Springer, 2016.

[116] Li-Jia Li, Hao Su, Li Fei-Fei, and Eric P Xing. Object bank: A high-level image representation for scene classification & semantic feature sparsification. In *Advances in neural information processing systems*, pages 1378–1386, 2010.

[117] Henry W Lim, Scott AB Collins, Jack S Resneck, Jean L Bolognia, Julie A Hodge, Thomas A Rohrer, Marta J Van Beek, David J Margolis, Arthur J Sober, Martin A Weinstock, et al. The burden of skin disease in the united states. *Journal of the American Academy of Dermatology*, 76(5):958–972, 2017.

[118] Dahua Lin and Xiaoou Tang. Recognize high resolution faces: From macrocosm to microcosm. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 1355–1362, 2006.

[119] Guangfeng Lin, Guoliang Fan, Liangjiang Yu, Xiaobing Kang, and Erhu Zhang. Heterogeneous structure fusion for target recognition in infrared imagery. In *Perception*

Beyond the Visible Spectrum Workshop, The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2015.

[120] Ce Liu, Jenny Yuen, and Antonio Torralba. Sift flow: Dense correspondence across scenes and its applications. *IEEE transactions on pattern analysis and machine intelligence*, 33(5):978–994, 2011.

[121] Guang-Hai Liu, Lei Zhang, Ying-Kun Hou, Zuo-Yong Li, and Jing-Yu Yang. Image retrieval based on multi-texton histogram. *Pattern Recognition*, 43(7):2380 – 2389, 2010.

[122] Jialu Liu, Chi Wang, Jing Gao, and Jiawei Han. Multi-view clustering via joint nonnegative matrix factorization. In *Proc. of SDM*, volume 13, pages 252–260. SIAM, 2013.

[123] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016.

[124] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[125] Fujun Luan, Sylvain Paris, Eli Shechtman, and Kavita Bala. Deep photo style transfer. *arXiv preprint arXiv:1703.07511*, 2017.

[126] Andrew L Maas, Awni Y Hannun, and Andrew Y Ng. Rectifier nonlinearities improve neural network acoustic models. In *Proc. ICML*, volume 30, 2013.

[127] S Madan, K J Dana, and G O Cula. Multimodal and time-lapse skin registration. *Skin Research And Technology: Official Journal Of International Society For Bioengineering And The Skin (ISBS) [And] International Society For Digital Imaging Of Skin (ISDIS) [And] International Society For Skin Imaging (ISSI)*, 2014.

[128] S.K. Madan, K.J. Dana, and O. Cula. Learning-based detection of acne-like regions using time-lapse features. In *Signal Processing in Medicine and Biology Symposium (SPMB), 2011 IEEE*, pages 1–6, Dec. 2011.

[129] S.C. Madeira and A.L. Oliveira. Biclustering algorithms for biological data analysis: a survey. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, 1(1):24–45, Jan 2004.

[130] I. Maglogiannis and C.N. Doukas. Overview of advanced computer vision systems for skin lesions characterization. *Information Technology in Biomedicine, IEEE Transactions on*, 13(5):721–733, Sept 2009.

[131] S. Maji, L. Bourdev, and J. Malik. Action recognition from a distributed representation of pose and appearance. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 3177–3184, June 2011.

[132] Stephen R Marschner, Stephen H Westin, Eric PF Lafortune, Kenneth E Torrance, and Donald P Greenberg. Image-based brdf measurement including human skin. In *Rendering Techniques 99*, pages 131–144. Springer, 1999.

[133] Alban Mathieu, Tom O Delmont, Timothy M Vogel, Patrick Robe, Renaud Nalin, and Pascal Simonet. Life on human surfaces: skin metagenomics. *Plos One*, 8(6):e65288, 2013.

[134] Alban Mathieu, Timothy Vogel, and Pascal Simonet. The future of skin metagenomics. *Research in Microbiology*, 165(2):69 – 76, 2014.

[135] The Mathworks, Inc., Natick, Massachusetts. *MATLAB version 8.5 (R2015a) and Neural Network Toolbox*, 2015.

[136] MATLAB Neural Network Toolbox. *version 8.3.0 (R2014a)*. The MathWorks Inc., Natick, Massachusetts, 2014.

[137] Elinor McKone, Nancy Kanwisher, and Bradley C Duchaine. Can generic expertise explain special processing for faces? *Trends in cognitive sciences*, 11(1):8–15, 2007.

[138] Kukizo Miyamoto, Hitomi Nagasawa, Yasuko Inoue, Kenichi Nakaoka, Ayaka Hirano, and Akira Kawada. Development of new in vivo imaging methodology and system for the rapid and quantitative evaluation of the visual appearance of facial skin firmness. *Skin Research & Technology*, 19(1):e525 – e531, 2013.

[139] William Montagna. *The structure and function of skin*. Elsevier, 2012.

[140] Andrew Y Ng, Michael I Jordan, Yair Weiss, et al. On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems*, 2:849–856, 2002.

[141] JuanCarlos Niebles, Chih-Wei Chen, and Li Fei-Fei. Modeling temporal structure of decomposable motion segments for activity classification. In Kostas Daniilidis, Petros Maragos, and Nikos Paragios, editors, *Computer Vision ECCV 2010*, volume 6312 of *Lecture Notes in Computer Science*, pages 392–405. Springer Berlin Heidelberg, 2010.

[142] Schommer Nina N. and Gallo Richard L. Review: Structure and function of the human skin microbiome. *Trends in Microbiology*, 21:660 – 668, 2013.

[143] Robert A Norman. The future of dermatological therapy and preventive dermatology. In *Preventive Dermatology*, pages 57–60. Springer, 2011.

[144] Maxime Oquab, Léon Bottou, Ivan Laptev, and Josef Sivic. Is object localization for free?-weakly-supervised learning with convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 685–694, 2015.

[145] Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. In *BMVC*, volume 1, page 6, 2015.

[146] G. Patterson and J. Hays. Sun attribute database: Discovering, annotating, and recognizing scene attributes. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2751–2758, June 2012.

[147] Scott B. Phillips, Nikiforos Kollias, Robert Gillies, Joseph A. Muccini, and Lynn A. Drake. Polarized light photography enhances visualization of inflammatory lesions of acne vulgaris. *Journal of the American Academy of Dermatology*, 37(6):948 – 952, 1997.

[148] L. Pishchulin, A. Jain, M. Andriluka, T. Thormahlen, and B. Schiele. Articulated people detection and pose estimation: Reshaping the future. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 3178–3185, June 2012.

[149] F. Porikli. Integral histogram: a fast way to extract histograms in cartesian spaces. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 829–836 vol. 1, June 2005.

[150] Javier Portilla and Eero P Simoncelli. A parametric texture model based on joint statistics of complex wavelet coefficients. *International journal of computer vision*, 40(1):49–70, 2000.

[151] Anderson R. Polarized light examination and photography of the skin. *Archives of Dermatology*, 127(7):1000–1005, 1991.

[152] Anthony P Raphael, Timothy A Kelf, Elizabeth MT Wurm, Andrei V Zvyagin, Hans Peter Soyer, and Tarl W Prow. Computational characterization of reflectance confocal microscopy features reveals potential for automated photoageing assessment. *Experimental dermatology*, 22(7):458–463, 2013.

[153] Jack Resneck and Alexa B Kimball. The dermatology workforce shortage. *Journal of the American Academy of Dermatology*, 50(1):50–54, 2004.

[154] Jack Resneck Jr. Too few or too many dermatologists?: Difficulties in assessing optimal workforce size. *Archives of dermatology*, 137(10):1295–1301, 2001.

[155] Gerald P Rodnan, Esther Lipinski, and Joan Luksick. Skin thickness and collagen content in progressive systemic sclerosis and localized scleroderma. *Arthritis & Rheumatology*, 22(2):130–140, 1979.

[156] Mariana Rosenthal, Deborah Goldberg, Allison Aiello, Elaine Larson, and Betsy Foxman. Skin microbiota: Microbial community structure and its potential association with health and disease. *Infection, Genetics & Evolution*, 11(5):839 – 848, 2011.

[157] 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 (IJCV)*, 115(3):211–252, 2015.

[158] Sreemanananth Sadanand and Jason J Corso. Action bank: A high-level representation of activity in video. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 1234–1241. IEEE, 2012.

[159] A. Sadovnik, A. Gallagher, D. Parikh, and Tsuhan Chen. Spoken attributes: Mixing binary and relative attributes to say the right thing. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 2160–2167, Dec 2013.

[160] Joshua D Safer. Thyroid hormone action on skin. *Dermato-endocrinology*, 3(3):211–215, 2011.

[161] Juan Luis Santiago Sánchez-Mateos, Carmen Moreno García del Real, Pedro Jaén Olasolo, and Salvador González. Reflectance-mode confocal microscopy in dermatological oncology. *Lasers in Dermatology and Medicine*, pages 285–308, 2011.

[162] Jason M Saragih, Simon Lucey, and Jeffrey F Cohn. Face alignment through subspace constrained mean-shifts. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 1034–1041. Ieee, 2009.

[163] Tiffany C Scharschmidt and Michael A Fischbach. What lives on our skin: Ecology, genomics and therapeutic opportunities of the skin microbiome. *Drug Discovery Today. Disease Mechanisms*, 10(3-4), 2013.

[164] Cordelia Schmid. Constructing models for content-based image retrieval. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, volume 2, pages II–39. IEEE, 2001.

[165] Divya Seth, Khatiya Cheldize, Danielle Brown, and Esther E Freeman. Global burden of skin disease: Inequities and innovations. *Current Dermatology Reports*, pages 1–7, 2017.

[166] Abhishek Sharma, Abhishek Kumar, Hal Daume III, and David W Jacobs. Generalized multiview analysis: A discriminative latent space. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2160–2167. IEEE, 2012.

[167] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(8):888–905, 2000.

[168] YiChang Shih, Sylvain Paris, Connelly Barnes, William T Freeman, and Frédo Durand. Style transfer for headshot portraits. 2014.

[169] Jamie Shotton, John Winn, Carsten Rother, and Antonio Criminisi. Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context. *International Journal of Computer Vision*, 81(1):2–23, 2009.

[170] M. Silveira, J.C. Nascimento, J.S. Marques, A. R S Marcal, T. Mendonca, S. Yamauchi, J. Maeda, and J. Rozeira. Comparison of segmentation methods for melanoma diagnosis in dermoscopy images. *Selected Topics in Signal Processing, IEEE Journal of*, 3(1):35–45, Feb 2009.

[171] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[172] Pawan Sinha, Benjamin Balas, Yuri Ostrovsky, and Richard Russell. Face recognition by humans: Nineteen results all computer vision researchers should know about. *Proceedings of the IEEE*, 94(11):1948–1962, 2006.

[173] Josef Sivic and Andrew Zisserman. Video google: A text retrieval approach to object matching in videos. In *null*, page 1470. IEEE, 2003.

[174] Eduardo Somoza, Gabriela Oana Cula, Catherine Correa, and Julie B Hirsch. Automatic localization of skin layers in reflectance confocal microscopy. In *Image Analysis and Recognition*, pages 141–150. Springer, 2014.

[175] Nitish Srivastava, Geoffrey E Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[176] Jiangwen Sun, Jinbo Bi, and Henry R Kranzler. Multi-view singular value decomposition for disease subtyping and genetic associations. *BMC genetics*, 15(1):73, 2014.

[177] Liang Sun, Shuiwang Ji, and Jieping Ye. Canonical correlation analysis for multilabel classification: A least-squares formulation, extensions, and analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(1):194–200, Jan 2011.

[178] Shiliang Sun. A survey of multi-view machine learning. *Neural Computing and Applications*, 23(7-8):2031–2038, 2013.

[179] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2015.

[180] Ruth E. Turnbaugh, Peter J.and Ley, Micah Hamady, Claire M. Fraser-Liggett, Rob Knight, and Jeffrey I. Gordon. The human microbiome project. *Nature*, 449:804–810, October 2007.

[181] Martina Ulrich and Susanne Lange-Asschenfeldt. In vivo confocal microscopy in dermatology: from research to clinical application. *Journal of biomedical optics*, 18(6):061212–061212, 2013.

[182] Dmitry Ulyanov, Andrea Vedaldi, and Victor S. Lempitsky. Instance normalization: The missing ingredient for fast stylization. *CoRR*, abs/1607.08022, 2016.

[183] M. Varma and A. Zisserman. Texture classification: are filter banks necessary? In *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, volume 2, pages II–691–8 vol.2, 2003.

[184] M. Varma and A. Zisserman. A statistical approach to texture classification from single images. *International Journal of Computer Vision: Special Issue on Texture Analysis and Synthesis*, 62(1–2):61–81, April 2005.

[185] A. Vedaldi and B. Fulkerson. VLFeat: An open and portable library of computer vision algorithms. Web page, 2008.

[186] Andrea Vedaldi and Karel Lenc. Matconvnet: Convolutional neural networks for matlab. In *Proceedings of the 23rd Annual ACM Conference on Multimedia Conference*, pages 689–692. ACM, 2015.

[187] Heng Wang, Alexander Klser, Cordelia Schmid, and Cheng-Lin Liu. Dense trajectories and motion boundary descriptors for action recognition. *International Journal of Computer Vision*, 103(1):60–79, 2013.

[188] Hongxing Wang, Chaoqun Weng, and Junsong Yuan. Multi-feature spectral clustering with minimax optimization. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 4106–4113. IEEE, 2014.

[189] Jianzhong Wang. Classical multidimensional scaling. In *Geometric Structure of High-Dimensional Data and Dimensionality Reduction*, pages 115–129. Springer Berlin Heidelberg, 2011.

[190] Kaiye Wang, Ran He, Wei Wang, Liang Wang, and Tieniu Tan. Learning coupled feature spaces for cross-modal matching. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 2088–2095, Dec 2013.

[191] Li-Yi Wei and Marc Levoy. Fast texture synthesis using tree-structured vector quantization. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 479–488. ACM Press/Addison-Wesley Publishing Co., 2000.

[192] Kilian Q Weinberger and Lawrence K Saul. Distance metric learning for large margin nearest neighbor classification. *The Journal of Machine Learning Research*, 10:207–244, 2009.

[193] Tim Weyrich, Wojciech Matusik, Hanspeter Pfister, Bernd Bickel, Craig Donner, Chien Tu, Janet Mcandless, Jinho Lee, Addy Ngan, Henrik Wann, and Jensen Markus Gross. Analysis of human faces using a measurement-based skin reflectance model. *ACM Transactions on Graphics*, 25:1013–1024, 2006.

[194] Jianxin Wu and J.M. Rehg. Beyond the euclidean distance: Creating effective visual codebooks using the histogram intersection kernel. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 630–637, 2009.

[195] Xinxiao Wu, Dong Xu, Lixin Duan, and Jiebo Luo. Action recognition using context and appearance distribution features. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 489–496, June 2011.

[196] Chang Xu, Dacheng Tao, and Chao Xu. A survey on multi-view learning. *arXiv preprint arXiv:1304.5634*, 2013.

[197] Yi Yang and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1385–1392, June 2011.

[198] PLJM Zeeuwen, M Kleerebezem, HM Timmerman, and J Schalkwijk. Microbiome and skin diseases. *CURRENT OPINION IN ALLERGY AND CLINICAL IMMUNOLOGY*, 13(5):514 – 520, 2013.

[199] Hang Zhang and Kristin Dana. Multi-style generative network for real-time transfer. *arXiv preprint arXiv:1703.06953*, 2017.

[200] Yimeng Zhang, Zhaoyin Jia, and Tsuhan Chen. Image retrieval with geometry-preserving visual phrases. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 809–816, June 2011.

[201] S. C. Zhu, Y. N. Wu, and D. Mumford. Filters, random field and maximum entropy: Towards a unified theory for texture modeling. *International Journal of Computer Vision*, 27(2):1–20, March/April 1998.