tvst

Article

Meibography Phenotyping and Classification From Unsupervised Discriminative Feature Learning

Chun-Hsiao Yeh[1,2], Stella X. Yu[1,3], and Meng C. Lin[2,3]

[1] International Computer Science Institute, Berkeley, CA, USA
[2] Clinical Research Center, School of Optometry, University of California, Berkeley, CA, USA
[3] Vision Science Group, University of California, Berkeley, CA, USA

Correspondence: Meng C. Lin, University of California, Berkeley, School of Optometry, 360 Minor Hall, Berkeley, CA 94720-2020, USA.
e-mail: mlin@berkeley.edu

Received: May 21, 2020
Accepted: October 19, 2020
Published: February 8, 2021

Keywords: meibography; meibomian gland dysfunction; atrophy; unsupervised feature learning; feature hierarchical clustering

Citation: Yeh C-H, Yu SX, Lin MC. Meibography phenotyping and classification from unsupervised discriminative feature learning. Trans Vis Sci Tech. 2021;10(2):4, https://doi.org/10.1167/tvst.10.2.4

Purpose: The purpose of this study was to develop an unsupervised feature learning approach that automatically measures Meibomian gland (MG) atrophy severity from meibography images and discovers subtle relationships between meibography images according to visual similarity.

Methods: One of the latest unsupervised learning approaches is to apply feature learning based on nonparametric instance discrimination (NPID), a convolutional neural network (CNN) backbone model trained to encode meibography images into 128-dimensional feature vectors. The network aims to learn a similarity metric across all instances (e.g. meibography images) and groups visually similar instances together. A total of 706 meibography images with corresponding meiboscores were collected and annotated for the use of network learning and performance evaluation.

Results: Four hundred ninety-seven meibography images were used for network learning and tuning, whereas the remaining 209 images were used for network model evaluations. The proposed nonparametric instance discrimination approach achieved 80.9% meiboscore grading accuracy on average, outperforming the clinical team by 25.9%. Additionally, a 3D feature visualization and agglomerative hierarchical clustering algorithms were used to discover the relationship between meibography images.

Conclusions: The proposed NPID approach automatically analyses MG atrophy severity from meibography images without prior image annotations, and categorizes the gland characteristics through hierarchical clustering. This method provides quantitative information on the MG atrophy severity based on the analysis of phenotypes.

Translational Relevance: The study presents a Meibomian gland atrophy evaluation method for meibography images based on unsupervised learning. This method may be used to aid diagnosis and management of Meibomian gland dysfunction without prior image annotations, which require time and resources.

Introduction

Meibomian gland dysfunction (MGD) is the most common underlying cause of dry eye syndrome where Meibomian glands (MGs) do not secrete enough lipids into the tears. The transillumination and infrared light are used to appreciate MG characteristics (i.e. measuring the percent of MG atrophy defined as the ratio of MG loss area to the total tarsal plate area) for MGD diagnosis.[1,2] Standardized MG atrophy grading scales have been developed to assess the severity of MG atrophy.[3,4]

In recent years, artificial intelligence (AI) in computer vision has arisen with deep convolutional neural networks (CNNs), which learned predicted features via supervised learning on a large dataset of labeled images.[5,6] AI has shown huge progress in the field of medicine, including cancer diagnosis, lung segmentation, and tumor detection,[7–9] especially in the ophthalmic domain. For example, AI has been applied to build models to detect subclinical Keratoconus,[10,11] which is the leading cause of corneal transplantation. Different AI systems were developed to detect the cases of glaucoma and have achieved promising performance.[12,13] AI has also benefited the MG

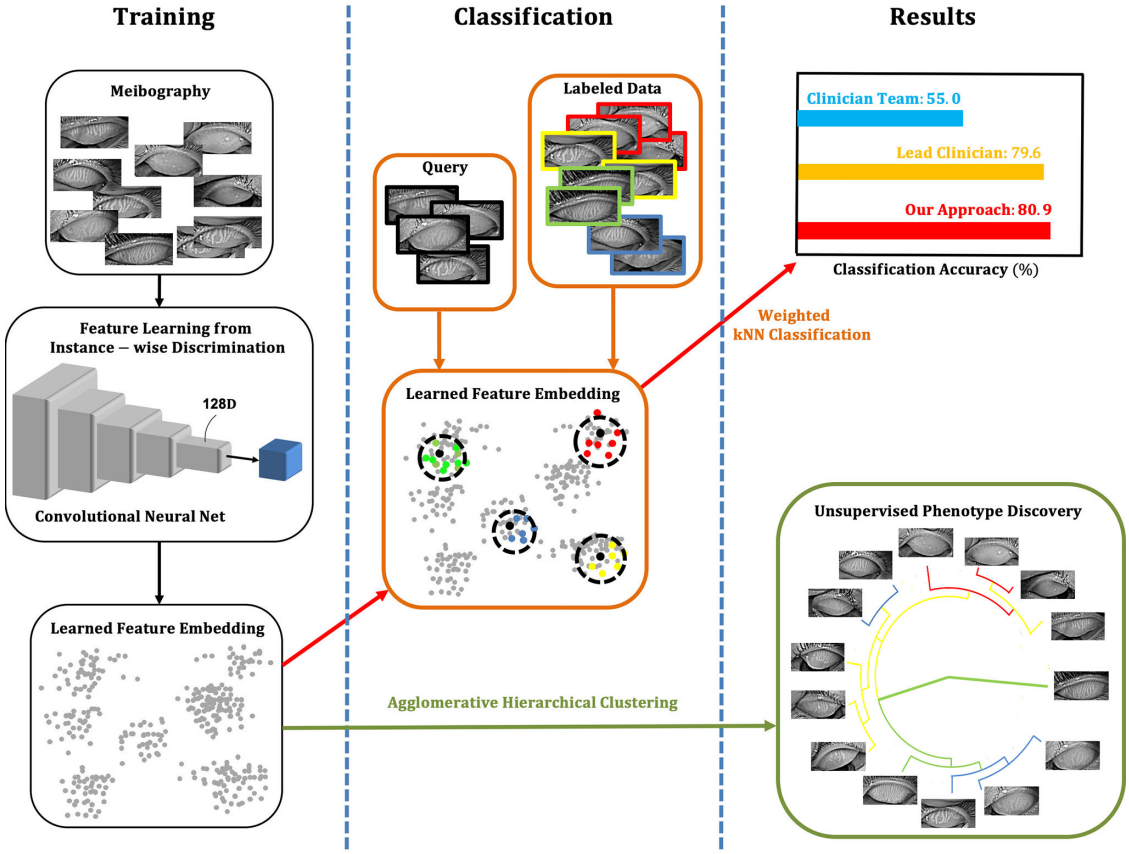

Figure 1. Overview of the approach. The nonparametric instance discrimination (NPID) is applied to learn a metric by feeding unlabeled meibography images, then to discriminate them according to their visual similarity. This approach aims for both measuring the atrophy severity and discovering subtle relationships between meibography images. There is no required image labelling, serving as ground truth for training.

atrophy evaluation from meibography images and have shown significantly improved performance.[14] However, it is costly, or sometimes even impossible for training CNNs on large labeled data sets because most of them have imbalanced label classes (i.e. one class accounts for almost 90% of the data, whereas other classes have far fewer samples). Additionally, vision data sets may contain labeling errors, leading to training issues for CNN models, especially for the class with a few samples.

Unsupervised representation learning aims to learn a robust embedding space from data without human annotation. Recently, discriminative approaches especially contrastive learning-based approaches, such as (nonparametric instance discrimination [NPID],[15] MoCo,[16] SimCLR,[17] etc) have gained most ground and achieved the state-of-the-art on standard large-scale image classification benchmarks with increasingly more computation and data augmentations. Based on our experience from extensive experimentation (cross-level discrimination [CLD][18]), NPID remains competitive, especially on small data sets.

Furthermore, some unsupervised methods could be extended to the semisupervised learning (i.e. LLP,[19] and CPC version 2),[20] by first learning in an unsupervised way and then fine-tuning with few labeled data. Note that more details are provided in the Discussion section.

In this paper, NPID[15] was applied for image analysis of MG from meibography to investigate MG features based on visual phenotypes. Furthermore, the visualization and hierarchical clustering algorithms were applied to show the feature clustering of meibography images. Whereas completely ignoring class labels, this unsupervised network discriminates between individual instances (e.g. meibography images) and automatically learns the similarity between instances, as shown in Figure 1. This approach automatically measures MG atrophy severity from meibography images, as well as discovers subtle relationships between

Table 1. Subject Demographics and Meiboscores of the Meibography Image Data Sets

	Train	Validation	Test
Images, *N*	398	99	209
Patient demographics			
Unique individuals, *N*	308	77	191
Age, average ± SD	25.5 ± 10.9	27.0 ± 12.6	26.4 ± 11.6
Female/total patients, %	63.5	66.6	68.3
Atrophy severity distribution, *n* (%)			
Meiboscore 0	73 (18.3)	18 (18.2)	38 (18.2)
Meiboscore 1	267 (67.1)	67 (67.7)	142 (67.9)
Meiboscore 2	53 (13.3)	13 (13.1)	27 (12.9)
Meiboscore 3	5 (1.3)	1 (1.0)	2 (1.0)

meibography images according to visual similarity. Additionally, an extensive experimental design was implemented to assess performance of evaluating MG atrophy by comparing the results obtained by the unsupervised learning method with those from a team of clinicians as well as a supervised learning method.

Method

Development and Test Dataset

Based on a previous study,[14] University of California, Berkeley Clinical Research Center recruited adult human subjects for a single-visit ocular surface evaluation, which included MG imaging for gland atrophy assessment, during the period from 2012 to 2017. Clinicians used the OCULUS Keratograph 5M (OCULUS, Arlington, WA), a clinical instrument that uses infrared light with wavelength 880 nm for MG imaging[21] to capture MG images of patients' upper and lower eyelids for both eyes. In this study, only upper eyelid images were used. A total of 706 images were collected after prescreening to rule out images that did not capture the entire upper eyelid. Each examining clinician assigned an MG atrophy severity score during the examination, namely the meiboscore. A previously published clinical grading criterion[3] was applied to define the MG percent atrophy and corresponding meiboscores. For example, the percent MG atrophy 0% is regarded as meiboscore 0, less than 33% as meiboscore 1, less than 66% as meiboscore 2, and the percent atrophy higher than 66% as meiboscore 3. The meiboscores were assigned by trained clinicians and were referred to as "clinical meiboscore." The subject demographics are shown in Table 1. Some samples of meibography image with corresponding meiboscores are shown in Figure 2.

Nonparametric Instance Discrimination

Figure 3 shows the overall pipeline for the proposed NPID approach. A standard CNN was utilized to form a feature vector through each image embedding, which was then normalized with Euclidean norm (L2-norm) to avoid overfitting and passed to a nonparametric softmax classifier for discriminating instances. The concept of attention layer and mask[22] were applied to form a scalar matrix representing the relative importance of layer activations at different 2D spatial locations with respect to the target task. The feature embedding was trained to learn a similarity metric across all instances and group visually similar instances closer together. This approach does not rely on image annotation, enabling efficient applications on real-world datasets without time-consuming labelling. It therefore scaled well to large data sets and deeper networks by using noise-contrastive estimation (NCE) to handle the computation cost that other approaches struggled with.

Nonparametric Instance Discrimination

Traditionally, most real-world applications (e.g. animal and car detection) can be developed by providing labeled data, which reduces it down to a classification problem. However, for tasks like MG atrophy evaluation of meibography images, such labeled data are not easily generated. The image annotation-related issues can be solved by learning a feature embedding function $f: X \mapsto R^d$, which maps images to a feature space of dimension d. The aim was to construct the

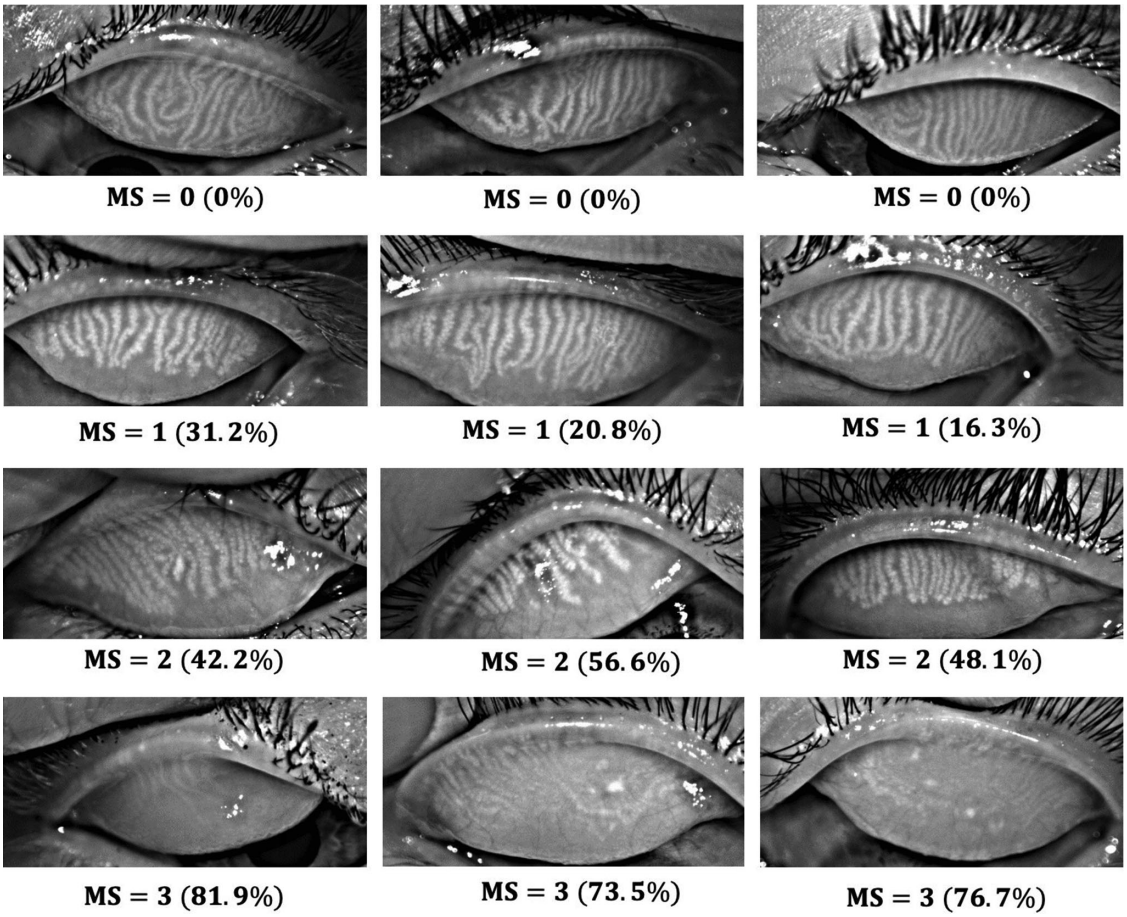

Figure 2. Meibography images with ground-truth percent atrophy (%) and ground-truth meiboscore (MS). Given a meibography image, the area of gland atrophy and eyelid are compared to estimate the percent atrophy, and are then converted to meiboscore based on the criteria in Table 1.

feature embedding in such a way that similar images ended up close to each other.

The feature embedding was constructed from a convolutional neural network, f_θ parameterized by θ. To achieve the desired property of having similar images close to each other, NPID was adopted to train the network, according to the previous work by Wu et al.[18] Each image in the training data set X was considered to be a distinct class and the feature outputs of the network were used to differentiate between image instances.

The model was trained using a nonparametric softmax, rather than a more traditional parametric version, on the output features. The probability of an image x belonging to the *i*: *th* class was then given by:

$$P(i|v) = \frac{exp\left(f_\theta\left(x_i\right) f_\theta\left(x\right)/\tau\right)}{\sum_{j=1}^{n} exp\left(f_\theta\left(x_j\right) f_\theta\left(x\right)/\tau\right)}, \qquad (1)$$

where τ is the parameter to control the density of the data distribution. The learning objective was given by minimizing the log-likelihood:

$$arg - \sum_{i=1}^{n} logP(i|f_\theta\left(x_i\right)), \qquad (2)$$

The training loss could interpret how far each $f_\theta(x_i)$ was formed from all other feature vectors. The approach aimed to minimize the log-likelihood in order to force the $f_\theta(x_j)$, which activated the same convolutional filters to be located in such same area in unit 128 dims hypersphere. During the learning process, all network parameter θ and the feature vector $f_\theta(x_i)$ were updated via stochastic gradient descent (SGD).[23]

Weighted KNN Classification

To classify an instance, $\hat{x}$ denoted in the validation set with the feature $\hat{v} = f_\theta(\hat{x})$ was computed and compared with all of the feature vectors $f_\theta(x_i)$ using cosine similarity: $f_\theta(x_i)f_\theta(\hat{x})$. The top k nearest neighbors N_k was then used to predict the class of $\hat{x}$ via

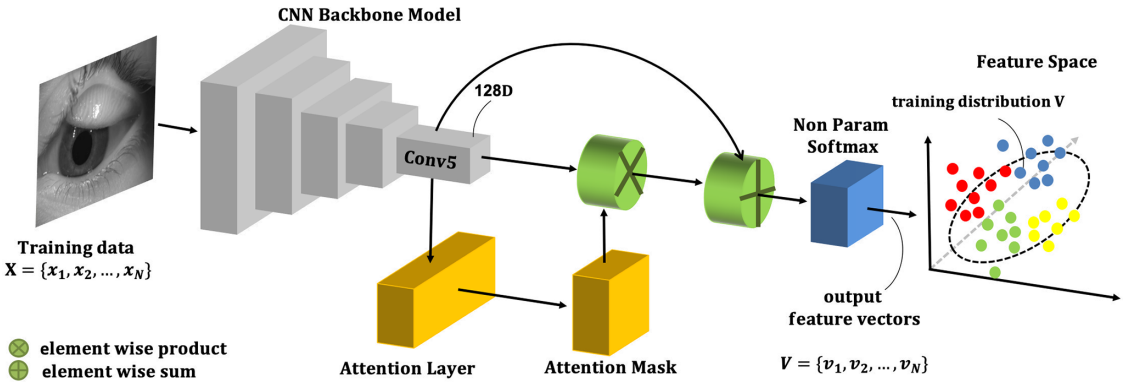

Figure 3. The pipeline for the nonparametric instance discrimination (NPID). A CNN backbone model, which encodes meibography images into 128-dimensional feature vectors during the learning procedure. The network aims to learn a similarity metric across all instances and grouped visually similar instances together. The attention layer and mask are applied to make the network model focus on significant parts of meibography image.

weighted voting. The class c obtained a total weight:

$$w_c = \sum_{i \in N_k} exp\left(\frac{f_\theta(x_i) f_\theta(\hat{x})}{\tau}\right) \cdot 1(c_i = c), \quad (3)$$

Here, $exp\left(\frac{f_\theta(x_i) f_\theta(\hat{x})}{\tau}\right)$ contributes to the weight of neighbor x_i, depending on cosine similarity. Note that $\tau = 0.07$ was chosen during network learning to carefully assess for picking the optimal k via the validation dataset (i.e. the best performance of NPID over the validation set was with $k = 25$). We follow the unsupervised as well as self-supervised representation learning literatures,[15–18,24] where cosine similarity has been used as a metric to describe the distance between two features on a unit sphere space.

Experiment

Experiments were extensively conducted to demonstrate the performance of the NPID approach. In the first experiment, the NPID network model with different structures, learning processes, techniques were evaluated. For the second experiment, the NPID network was compared against the performance of clinical grading and supervised learning algorithm.

Experimental Protocol

It is essential to evaluate the performance of the learned network model. The model was first evaluated on the validation set to select the hyperparameters that achieved the best performance. After fixing the optimal hyperparameters performed on the validation set, further evaluation was performed on the test set.

As illustrated in Figure 4, when the adapted threshold was set to be 0.25%, the classifications for the images with percent atrophy of 0% to 0.25%, 32.75% to 33.25%, and 65.75% to 66.25% remained ambiguous. For further clarification, labels of meibography were defined by using the dot plot as illustrated in Figure 4. The color of the central dot point refers to the ground-truth label, whereas the color of the outline refers to the appended label after the applied adapted threshold.

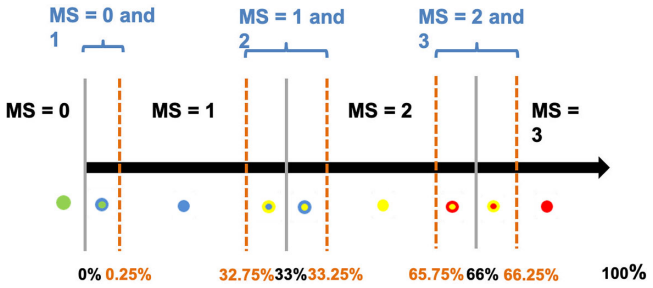

Figure 4. Relaxed meiboscore conversion rule with the adapted threshold. The percent atrophy to the meiboscore conversion criteria is relaxed with an adapted threshold near the grading transition limits (0%, 33%, and 66%). The threshold is set to be 0.25%. Percent atrophy falls in 0% to 0.25%, 32.75% to 33.25%, or 65.75% to 66.25% is acceptable to have both its ground-truth and adjacent meiboscores as correct prediction. The colors of the central dots refer to the ground-truth labels, whereas the colors of the outlines refer to the appended labels after applied the adapted threshold, which relaxes the criteria.

Because the ground-truth meiboscores were obtained from converting the percent atrophy of annotated meibography images using the conversion criteria (i.e. 0–33% = Meiboscore 1; 33–66% = Meibosore 2; and > 66% = Meibosore 3), the meibography images near the grading transition limits (0%, 33%, and 66%) were visually similar and difficult to classify due to small differences. For instance, when

Table 2. Checklist of the NPID Approach with Different Network Model Structures, Learning Processes, Data Argumentations, and Evaluation Techniques (Three Protocols are Illustrated in the Following Experiment)

	ResNet 50	Data Argumentation	Adapted Threshold	Attention Mechanism
Protocol 1	√	√		
Protocol 2	√	√	√	
Protocol 3	√	√	√	√

Table 3. The Performance of NPID with Three Different Protocols

	Protocol 1		Protocol 2		Protocol 3	
Evaluation	Top 1 (%)	Top 5 (%)	Top 1 (%)	Top 5 (%)	Top 1 (%)	Top 5 (%)
250 epochs	42.7 ± 0.4	86.3 ± 1.2	56.1 ± 0.5	86.2 ± 2.2	66.6 ± 0.8	91.8 ± 2.1
350 epochs	47.1 ± 0.7	87.3 ± 1.5	57.3 ± 1.3	84.8 ± 1.4	**67.3 ± 1.6**	**93.3 ± 1.3**
400 epochs	36.9 ± 1.1	83.7 ± 1.6	52.1 ± 0.9	85.6 ± 1.8	65.2 ± 0.5	90.9 ± 1.7

The best performance (protocol 3) achieves the top-1 accuracy of 68.4% and the top-5 accuracy of 6% by adding the adapted threshold, the attention mechanism and the network model learned with 350 epochs. Noted that the top-1 accuracies are reported in average accuracy ± standard deviation.

two meibography images are with 32.9% and 33.1% atrophy, they could be classified as with meiboscore 1 or 2, respectively. Therefore, an adapted threshold was warranted to reduce classification errors as suggested previously.[15]

Network Training Details

ResNet 50[25] was adopted as backbone network, which encoded the output as 128-dimensional vectors in all of the experiments. The network was trained using SGD with momentum 0.9 with a batch size of 32 and set the weight decay hyperparameter to 4×10^5. Learning-rate drop policy was carefully adjusted to obtain the best performance of the network on the validation data set. Data-augmentation techniques were adapted to each meibography image: 400 × 400 pixels were randomly cropped out from a given meibography image with 420 × 420 pixels, while a center crop of 400 × 400 pixels was made to meibography images for both validation and test data sets.

Algorithm Performance

Tables 2 and 3 show different protocol setups and the performance of each protocol, respectively. ResNet 50 was used as a backbone CNN with an embedded 128 dimensions feature vector. As noted, $\tau = 0.07$ and $k = 25$, with an initial learning rate of 0.005 were used as parameters. The prevalent hyperparameter selection approach has been applied for unsupervised as well as self-supervised learning,[15,18] where the hyperparameters are selected according to the labeled data in the downstream classification task. It can also be selected according to some criteria, such as normalized mutual information or image retrieval accuracy in the unlabeled validation set.[18] The downstream classification performance is dependent on a particular data set and may be sensitive to tau. However, in the present study, we applied the same tau value of 0.07 for unsupervised representation learning over the ImageNet data set as reported in the published code.[15]

The network model was trained from scratch without using any pretrained model in this experiment. The best performance achieved top-1 = 68.4% and top-5 = 93.6% by adding the adapted threshold and the attention mechanism. In addition, the network model is learned with 350 epochs. The training stops when there is a convergence of loss (i.e. loss is not decreasing much or stabilizing). This implies that the correct number of the epoch has been achieved. The training of the network ResNet 50 with 400 epochs maximum reached a steady level after 400 epochs. The training was also stopped after 400 epochs in order to prevent overfitting.

The top-1 accuracy is noted as the conventional accuracy, referring to the expected model prediction for the label of the nearest neighbor in the feature space; whereas the top-5 accuracy refers to the model prediction taking the label of 5 closest neighbors as reference. It is also worth noting that an epoch was defined as an

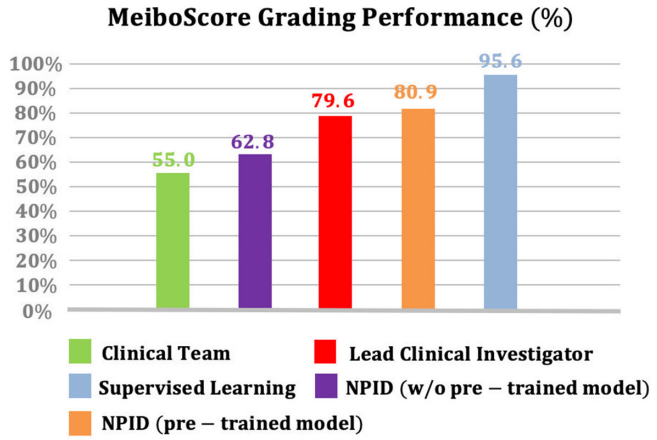

Figure 5. Meiboscore grading performance of clinicians and algorithm (%). The NPID approach was compared against the clinical team (clinical meiboscore),[14] the lead clinical investigator,[14] and a supervised learning approach.[14] The NPID approach achieves 80.9% overall grading accuracy with ImageNet pretrained model.

entire data set passing forward and backward through the neural network in one cycle.

Further investigation found that the best performance (protocol 3; see Table 3) achieved top-1 = 67.3 ± 1.6% and top-5 = 93.3 ± 1.3% by adding the adapted threshold and attention mechanism. The network model was learned with 350 epochs.

In Figure 5, the NPID approach was compared against the clinical team (clinical meiboscore), the lead clinical investigator (LCI), and a supervised learning approach.[14] The NPID approach achieved 80.9% overall grading accuracy with ImageNet pretrained model, which outperformed the clinical team grading by 25.9% and lead clinical investigator by 1.3%. The NPID approach accuracy without using ImageNet pretrained model was also provided to show that the pretrained model benefited the performance by gaining around 14% accuracy. The ImageNet pretrained model[26] was used on a large set of real-world images, providing a useful starting point for restoring a pretrained model. The ImageNet model already had the ability to adapt features from many tasks or different kinds of images. It is important to note that the ground-truth meiboscores were obtained from the percent MG atrophy, calculated from human-annotated segmentation masks.

Additional experiments were conducted to show the grading performance of the proposed method by each class and the instance average accuracy from 10 runs of each protocol (see Table 4). These results suggested that 200 epochs were needed to conduct a fair comparison with the supervised learning approach.[14]

T-test Analysis

To compare the 10 runs accuracies among different settings (LCI, clinical team, and our NPID), *t*-tests on the comparisons were performed. The entire data D (706 images) was divided into train, validation, and test sets according to 56% / 14% / 30%, respectively. Specifically, images were randomly picked from each meiboscore based on the meiboscore distribution (see Table 1 in the manuscript) to avoid data imbalance. This process was repeated for 10 times and the performance of NPID, LCI, and clinician team was evaluated over 10 different test sets.

Accuracies for NPID, LCI, and clinicians are reported in Table 5. The difference in performance was statistically significant only between NPID and clinicians.

Multiclass Classification

The K-class classification is to categorize the meibography (MG) data, which is graded by clinicians based on the atrophy severity (i.e. meiboscore 0 = none, 1 = mild, 2 = moderate, and 3 = severe). For 2-class classification, none and mild MG data are categorized together, whereas moderate and severe MG data are grouped. For 3-class classification, only moderate and severe MG data are assigned to be in the same class, and contrast to none, mild MG data. The evaluation protocols are defined in Table 6.

Table 4. Meiboscore Grading Performance of the Proposed Algorithm (%) by Each Class and Instance Average Accuracy ± Standard Deviation Over 10 Runs

	NPID (w/o Pretrained Model) Top 1 (%)	NPID (Pretrained Model) Top 1 (%)
Meiboscore 0	58.0 ± 0.8	71.1 ± 1.1
Meiboscore 1	63.4 ± 1.1	82.4 ± 0.5
Meiboscore 2	74.0 ± 0.6	85.2 ± 1.7
Meiboscore 3	50.0 ± 0.0	50.0 ± 0.0
Instance avg.	63.6 ± 2.3	**80.4 ± 2.1**

Table 5. The Accuracies (%) for Our NPID, LCI, and Clinicians

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	Confidence Interval (CI)	*P* Value (NPID vs. LCI or Clinician)
NPID	74.1	76.5	73.9	81.8	82.1	72.6	83.3	79.8	81.7	80.4	[75.8 – 81.4]	
LCI	73.1	77.4	72.4	78.9	76.6	74.4	80.7	76.8	81.9	82.4	[75.0 – 80.0]	0.498
Clinicians	53.4	54.5	52.1	57.9	52.8	58.6	59.3	56.8	58.7	56.4	[54.2 – 58.0]	<0.001

Note that P1 to P10 were referred to as our "10 random processes of data selection." The last column lists the associated *P* values from paired *t*-test between the row approach and NPID. The *P* value between LCI and the clinician team is < 0.001. These results demonstrate that our NPID is on par with LCI and significantly better than the clinician team.

Table 6. Evaluation Protocols of 2-, 3-, and 4-Class Classification

	2-Class Classification	3-Class Classification	4-Class Classification
Protocol Setting	[none, mild] vs. [moderate, severe]	[none] vs. [mild] vs. [moderate, severe]	[none] vs. [mild] vs. [moderate] vs. [severe]

Table 7. The Top-1 Average $\pm$ Standard Deviation Accuracy (%) for 2-, 3-, and 4-Class Classification by Each class of Meiboscore and the Class Average Accuracy

	2-Class	3-Class	4-Class
Class of meiboscore 0	92.7 ± 0.5	73.7 ± 1.2	71.1 ± 1.3
Class of meiboscore 1		80.9 ± 0.7	82.4 ± 0.7
Class of meiboscore 2	86.2 ± 1.8	89.7 ± 1.6	81.5 ± 1.8
Class of meiboscore 3			50.0 ± 0.0
Class avg. accuracy	89.5 ± 1.0	82.3 ± 1.2	71.3 ± 1.0
Instance avg. accuracy	85.2 ± 1.9	81.3 ± 2.1	80.8 ± 2.3

Table 7 reported the top-1 accuracy by each class of meiboscore and the class average accuracy. In the 4-class classification, most wrong predictions of none were classified as mild, and most wrong predictions of moderate were misclassified as severe.

By combining these similar atrophy severities (e.g. moderate and severe) into one superclass, our approach delivers better performance at coarse-grained categorization.

Feature Visualization

Figure 6 shows the 2D t-SNE[27] visualization of the proposed best feature embedding with ImageNet pretrained model (see Fig. 5). A total of 209 meibography images were used to pass through the network model and the feature was then squeezed from 128D to 2D. It is easy to find out that some types of meibography features were grouped closely or located in the same area in the unit hypersphere. For example, the meibography images with a yellow central dot outlined by red, and red dots are located in the same area in the unit hypersphere because they have visually similar phenotypes. Specifically, it is observed that most yellow dots are located closely in the center of the plot.

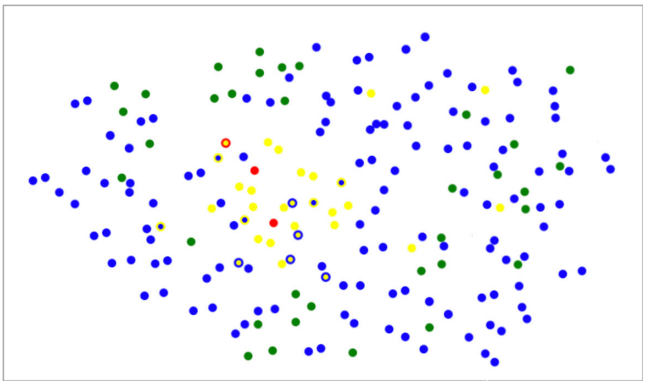

Figure 6. The 2D t-SNE visualization. A total of 209 meibography images in the test data are used to pass through the network model and the feature is collapsed from 128D to 2D. A color is designated to each feature dot of the meibography image based on the phenotypes listed in the manuscript.

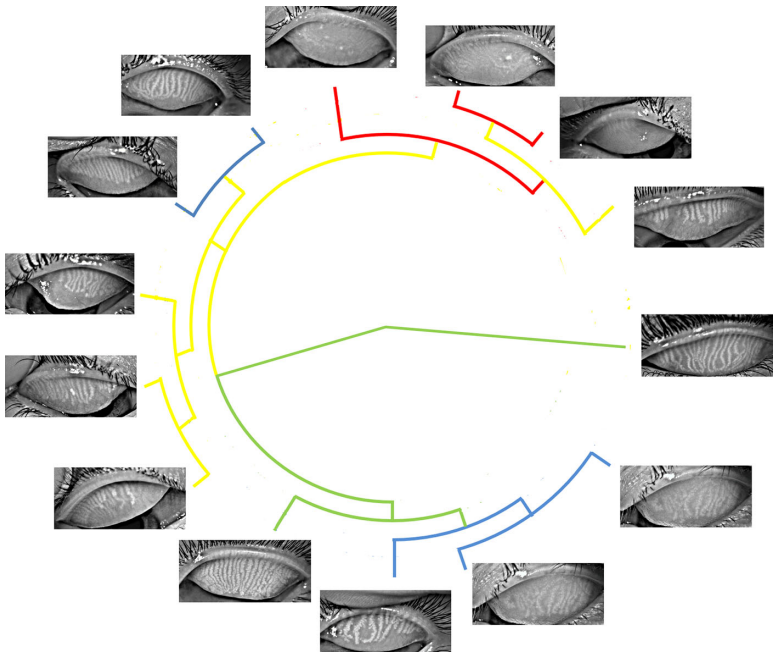

Figure 7. The agglomerative hierarchical clustering is an abstract representation of how images of meibography types are separated in the feature vector space. Note that the leaves (e.g. meibography images) represent feature vector centroids of 40 training images of each meibography phenotype (e.g. meiboscore 0, 1, 2, and 3).

Unsupervised Learning of Visual Hierarchies and Clustering

The agglomerative hierarchical clustering in Figure 7 was based on the generated feature by applying a clustering algorithm.[28] The leaves represent feature vector centroids of around 40 training images of each meibography phenotype (e.g. meiboscores 0, 1, 2, and 3). The agglomerative clustering tree is an abstract representation of how meibography image types are separated in the feature vector space. It also illustrates how visually similar the eight types (see Fig. 4) of meibography images are to the trained NPID model. By investigating the clustering tree, it is easy to see that meibography images with the blue (meiboscore of 1) and green dot (meiboscore of 0) were grouped together in the first stage, whereas images with meiboscore 2 and 3 images were connected. Hierarchical clustering from 14 clusters to 4 clusters was observed through the clustering tree.

Discussion

The present work develops an unsupervised feature learning approach that automatically measures MG atrophy severity from meibography images and discovers relationships between meibography images according to visual similarity. To the best of our knowledge, the proposed work is among the first to use unsupervised feature learning to measure MG atrophy severity, which is distinctive to many of other approaches (e.g. supervised learning) in evaluating MG atrophy.

Our experiments on the test data set (see Table 1) confirm the effectiveness of our framework and its superiority over clinical assessments. The proposed NPID approach achieved 80.9% meiboscore grading accuracy on average, outperforming the clinical team by 25.9% and the LCI by 1.3%. Additionally, another advantage is that a 3D feature visualization and an agglomerative hierarchical clustering algorithm are provided to discover subtle relationships between meibography images.

In future work, the proposed method could be extended to the semi-supervised learning by first learning from the big unlabeled data (706 images in our case) and then fine-tuning the network on a small fraction (e.g. 10% of the entire data set) of labeled data. LLP[19] suggests that such scenarios can benefit the unsupervised learning and can give an extra boost of performance. CPC version 2[20] shows representation learning can be used in semi-supervised learning schemes to drastically reduce the number of labeled images. TVOS also demonstrates the concept to videos.[29]

In real-world applications, learned features from deep learning methods via supervised learning on a

large annotated data have shown promising performance.[14] However, obtaining annotated information on the meibomian gland structure for network learning is time consuming.[2] The advantage of the proposed work (e.g. unsupervised discriminative feature learning) is to analyze the MG atrophy and potentially other features (future work) by incorporating the appropriate algorithms for analyzing raw and unprocessed images so that doctors could gain timely impression of MG features and prognosis of MGD immediately after image capture. This image analysis technology could also be applied to other ophthalmic conditions, such as keratoconus (KC) and glaucoma.[10,11] However, in the proposed work, other factors (e.g. age, gender, and race) are not considered for analyzing the MG atrophy. Therefore, future work can investigate how the discovered relationships between meibography images using the NPID approach may be influenced by demographic data and other ocular health-related information to further our understanding about the potential risk factors of MGD.

Acknowledgments

The authors thank Dorothy Ng, Jessica Vu, Jasper Cheng, Kristin Kiang, Megan Tsiu, Fozia KhanRam, April Myers, Shawn Tran, Michelle Hoang, and Zoya Razzak for providing annotations for the meibography images.

We are also grateful for general support from UCB-CRC Unrestricted Fund; Roberta J. Smith Research Fund.

Disclosure: **C.H. Yeh,** None; **S.X. Yu,** None; **M.C. Lin,** None

References

1. Arita R, Itoh K, Inoue K, Amano S. Noncontact infrared meibography to document age-related changes of the meibomian glands in a normal population. *Ophthalmology*. 2008;115:911–915.
2. Pult H, Nichols JJ. A review of meibography. *Optom Vis Sci*. 2012;89:E760–E769.
3. Pflugfelder SC, Tseng SCG, Sanabria O, et al. Evaluation of subjective assessments and objective diagnostic tests for diagnosing tear-film disorders known to cause ocular irritation. *Cornea*. 1998;17:38–56.
4. Arita R, Itoh K, Maeda S, et al. Proposed diagnostic criteria for obstructive meibomian gland dysfunction. *Ophthalmology*. 2009;116: 2058–2063.
5. Krizhevsky A, Sutskever I, Hinton GE.ImageNet classification with deep convolutional neural networks. 2012:1097–1105. Available at: http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neuralnetworks. Accessed February 8, 2019.
6. Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. International Conference on Medical Image Computing and Computer-Assisted Intervention. Switzerland AG: Springer; 2015: 234–241.
7. Liu Y, Gadepalli K, Norouzi M, et al. Detecting cancer metastases on gigapixel pathology images. 3 2017. Available at: http://arxiv.org/abs/1703.02442. Accessed February 8, 2019.
8. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist- level classification of skin cancer with deep neural networks. *Nature*. 2017;542:115–118.
9. Litjens G, Kooi T, Bejnordi BE, et al. A survey on deep learning in medical image analysis. *Med Image Anal*. 2017;42:60–88.
10. Cao K, Verspoor K, Sahebjada S, Baird PN. Evaluating the performance of various machine learning algorithms to detect subclinical keratoconus. *Trans. Vis. Sci. Tech*. 2020;9(2):24.
11. Velázquez-Blázquez JS, Bolarín JM, Cavas-Martínez F, Alió JL. EMKLAS: a new automatic scoring system for early and mild keratoconus detection. *Trans Vis Sci Tech*. 2020;9(2):30.
12. Russakoff DB, Mannil SS, Oakley JD, et al . A 3D deep learning system for detecting referable glaucoma using full OCT macular cube scans. *Trans Vis Sci Tech*. 2020;9(2):12.
13. Thompson AC, Jammal AA, Medeiros FA. A review of deep learning for screening, diagnosis, and detection of glaucoma progression. *Trans Vis Sci Tech*. 2020;9(2):42.
14. Wang J, Yeh TN, Chakraborty R, Yu SX, Lin MC. A deep learning approach for Meibomian gland atrophy evaluation in meibography images. *Trans Vis Sci Tech*. 2019;8(6):37.
15. Wu Z, Xiong Y, Yu SX, Lin D. Unsupervised feature learning via non-parametric instance discrimination. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018;3733–3742.
16. He K, Fan H, Wu Y, Xie S, Girshick R. Momentum contrast for unsupervised visual representa-

tion learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020. 9729–9738.
17. Chen T, Kornblith S, Norouzi M, Hinton G. A simple framework for contrastive learning of visual representations. In Proceedings of the 36th International Conference on Machine Learning (ICML 2020); 2020;119:1597–1607.
18. Wang X, Liu Z, Yu SX. Unsupervised feature learning by cross-level discrimination between instances and groups. arXiv preprint arXiv:2008.03813 2020.
19. Zhuang C, Ding X, Murli D, Yamins D. Local label propagation for large-scale semi-supervised learning. 2019. arXiv preprint arXiv:1905.11581.
20. Hénaff Olivier J, et al. Data-efficient image recognition with contrastive predictive coding. 2019. arXiv preprint arXiv:1905.09272.
21. Markoulli M, Duong TB, Lin M, Papas E. Imaging the tear film: a comparison between the subjective Keeler Tearscope-PlusTM and the Objective Oculust Keratograph 5M and Lipi- Viewt Interferometer. *Curr Eye Res*. 2018;43:155–162.
22. Jetley S, Lord NA, Lee N, Torr PHS. Learn to pay attention. 2018. arXiv preprint arXiv:1804.02391.
23. Robbins H, Monro S. A stochastic approximation method. *Ann Math Stat*. 1951;22(3):400–407.
24. Tian Y, Krishnan D, Isola P. Contrastive multiview coding. In Proceedings of the European Conference on Computer Vision (ECCV). arXiv: 1906.05849 2020.
25. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, arXiv: 2016;770–778.
26. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems. 2012;60(6):1097–1105.
27. Maaten L, Hinton G. Visualizing data using t-SNE. *J Machine Learn Res*. 2008;9(Nov):2579–2605.
28. Sarfraz S, Sharma V, Stiefelhagen R. Efficient parameter-free clustering using first neighbor relations. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. arXiv.org 2019: 8934–8943.
29. Zhang Y, Wu Z, Peng H, Lin S. A Transductive Approach for Video Object Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020:6949–6958.