Caring is Sharing – Exploiting the Value in Data for Health and Innovation
M. Hägglund et al. (Eds.)

doi:10.3233/SHTI230312

Few-Shot Meta-Learning for Recognizing Facial Phenotypes of Genetic Disorders

Ömer SÜMER[a,1], Fabio HELLMANN[a], Alexander HUSTINX[b], Tzung-Chien HSIEH[b], Elisabeth ANDRÉ[a] and Peter KRAWITZ[b]
[a] *Chair for Human-Centered Artificial Intelligence, University of Augsburg, Germany*
[b] *Institute for Genomic Statistics and Bioinformatics, University of Bonn, Germany*

Abstract. Computer vision has useful applications in precision medicine and recognizing facial phenotypes of genetic disorders is one of them. Many genetic disorders are known to affect faces' visual appearance and geometry. Automated classification and similarity retrieval aid physicians in decision-making to diagnose possible genetic conditions as early as possible. Previous work has addressed the problem as a classification problem; however, the sparse label distribution, having few labeled samples, and huge class imbalances across categories make representation learning and generalization harder. In this study, we used a facial recognition model trained on a large corpus of healthy individuals as a pre-task and transferred it to facial phenotype recognition. Furthermore, we created simple baselines of few-shot meta-learning methods to improve our base feature descriptor. Our quantitative results on GestaltMatcher Database (GMDB) show that our CNN baseline surpasses previous works, including GestaltMatcher, and few-shot meta-learning strategies improve retrieval performance in frequent and rare classes.

Keywords. Facial genetics, rare genetic disorders, image analysis, few-shot learning, meta-learning, imbalanced data, deep learning.

1. Introduction

Genetic disorders affect more than 5% of the population [1]; however, physicians might fail to spot and clinically diagnose most of them. There is a set of genetic conditions, and 30-40% of them are known to affect craniofacial development and facial morphology [2], and computer vision can help recognize skull alterations from facial images [3]. The output of such a system can support physicians in diagnosing rare syndromes and eventually lead to therapeutic interventions.

Previous literature uses geometric information, facial landmarks, and handcrafted features around face regions [4], however, a small number of subjects and syndromes limit their use in clinical settings. Shukla et al. [5] combined convolutional neural network features in face regions and used SVM classifiers. Recent studies [6,7] showed that end-to-end deep learning-based methods could substantially improve facial phenotyping.

Still, the number of samples in real-life situations and databases shows considerable variation across disorders. This makes training deep convolutional networks not feasible, as in any object classification task. The nature of the problem necessitates addressing

[1] Corresponding Author: Ömer Sümer, E-mail: oemer.suemer@informatik.uni-augsburg.de.

data imbalance and few-shot classification in facial phenotype analysis. Collecting facial images of rare facial genetic disorders requires lots of effort. Most of the previous works do not have publicly available databases to benchmark computer vision methodologies. GestaltMatcher Database[2] [7] is a recent effort to carry automated facial phenotyping forward. This paper presents a deep learning baseline that depends on a better facial recognition model and a few-shot meta-learning approach for unseen facial genetic disorders based on a highly imbalanced distribution of disorders.

2. Method

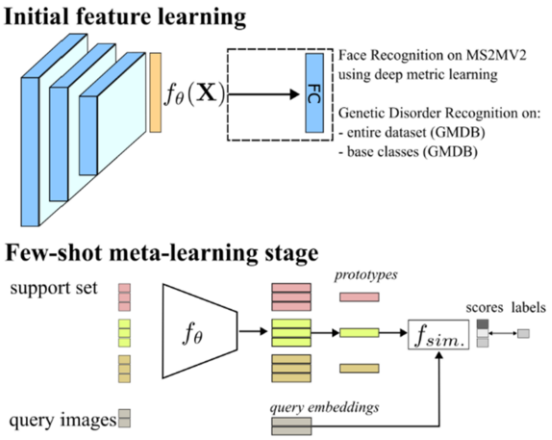

Figure 1. Workflow of initial feature learning and few-shot meta-learning: top) the initial feature learning is done either on face recognition task or genetic disorder classification; bottom) the learned representation is used in the few-shot meta-learning stage.

Figure 1 depicts the workflow of our proposed approach for facial phenotype recognition. The initial step in facial phenotype learning is to learn a solid initial representation. We trained a convolutional neural network backbone for this task by adopting the metric learning-based Arcface loss [8] in face recognition. Subsequently, the few-shot meta-learning stage aims to learn facial phenotypes from highly imbalanced data where most categories have limited samples.

There are separate support and query sets to learn to compare in the training and testing phases. These sets are created in an episodic manner. The bottom part of Figure 1 takes sampled episodes of support and query images, and first extracts features using the backbone encoder by initializing from face recognition pre-trained weights. The few-shot learning is defined according to the number of categories (N) and samples (K) in each support group in the support set. The task is described as K-way N-shot in previous literature [9]. During the training, the centroid of each embedding vector per class c is calculated: $c_k = \frac{1}{S_k}\sum_{x_i,y_i \in S_k} f_\Theta(x_i)$ where x_i and y_i are images and corresponding labels in each group of support set, S_k.

Furthermore, differing from [9], each episode has several K-way N-shot tasks. It refers to predicting the category of a query sample from K classes or N examples per

[2] It is accessible for clinicians and computer scientists under the following link: https://db.gestaltmatcher.org/

class in the support set. This setting learns a feature embedding that can retrieve samples belonging to the same category using a similarity metric. The main difference here is that meta-learning is independent of the tasks and can better generalize on unseen classes. In Prototypical Networks [9], the distance (or similarity) function is Euclidean distance. However, previous literature in facial phenotype recognition [7] used cosine similarity for the retrieval task. To make our meta-training as compatible as possible with our end task, we used cosine similarity between query embeddings, $f(x_i)$ and class centroids, c_k, and calculated logits as follows:

$$p(y = k \mid i) = \frac{e^{\tau \cdot \langle f_\Theta(x_i), c_k \rangle}}{\sum_k' e^{\tau \cdot \langle f_\Theta(x_i), c_k \rangle}} \quad (1)$$

where τ is a learnable scalar that we applied to scale the values before applying the Softmax function following the related literature [10,11].

We conducted our experiments on version 1.0.3 of the GMDB [7]. In version 1.0.3, the database contains 7.459 images of 449 syndromes. In addition to training and validation sets, there are two separate galleries and test sets for frequent and rare disorders. In both, faces are detected and aligned by RetinaFace [8]. Using five facial key points, we performed 5-point similarity alignment and normalized faces to the size of 112x112 pixels. During the training of baseline classification and few-shot meta-learning models, we only applied channel mean and standard deviation normalization according to the train set statistics and random horizontal flipping. When training both whole-set classification and few-shot meta-training, we used an SGD solver with a constant learning rate of 0.001 and weight decay of 0.0005 for 25 epochs. We used validation retrieval performance, specifically, the nearest neighbor retrieval of validation samples' feature embeddings to all training sets for model selection. During the few-shot meta-training, we sampled each episode containing four tasks, and the total number of episodes was kept at 100 and trained for 25 epochs.

Using the 512-dimensional embedding vector as feature representation, we evaluated the performance of our classification and nearest-neighbor approach in terms of top-k accuracies in the frequent and rare test sets in the GMDB. Following Hsieh et al. [7], learned facial embeddings were evaluated using two settings as follows:

1. The retrieval task reports top-k accuracies using k-nearest neighbors based on feature embeddings and cosine distances from the frequent gallery and frequent test sets.
2. The retrieval task reports top-k accuracies using k-nearest neighbors based on feature embeddings and cosine distances from the 10-Fold Cross-Validation rare gallery and rare test sets.

In all experiments, we calculate accuracies for Top-{1, 5, 10, 30} retrieval. We also reported the classification performance of [7] that reports only top-k accuracies using softmax outputs based on the frequent test set.

3. Results

Table 1 depicts the results of our ablation study. As we aim to improve the retrieval performance on both tasks, we only evaluated GestaltMatcher DCNN using predictions trained with cross-entropy loss. The performance of GestaltMatcher DCNN trained on v1.0.3 of the database is aligned with the published results in [7]. Top-1 accuracy varies in the ranges of 15% to 21% in frequent and rare sets where the total number of classes

is 204 and 245, respectively. Our stronger baseline, a ResNet-50 trained on MS1MV2 using ArcFace loss (Enc-healthy), performed 34.06% top-1 accuracy in the frequent set, whereas GestalthMatcher DCNN's retrieval performance remains at 15.96%.

Few-shot meta baseline that we adopted in our experiments is a 10-way 3-shot task with 2 query samples in each task. Following [11], we sampled multiple tasks in each episode. The reported experiments are done using four tasks per episode. Table 2 shows the retrieval performance of few-shot meta-learning models on both frequent and rare test sets.

Table 1. Performance comparison of GestaltMatcher DCNN and our baseline models on GMDB (v1.0.3).

Method	Top-1	Top-5	Top-10	Top-30
Frequent Set				
GestaltMatcher DCNN (7)				
Classification	21.21	42.08	54.60	73.92
Retrieval	15.96	33.83	45.46	69.64
Enc-healthy	34.06	53.96	64.42	**81.28**
Enc-all (GMDB)	**42.50**	58.18	65.26	78.08
Enc-base (GMDB)	40.47	**60.71**	**67.29**	79.09
Rare Set				
GestaltMatcher DCNN (7)				
Retrieval	19.26	36.28	44.07	60.73
Enc-healthy	26.31	42.62	46.98	62.92
Enc-all (GMDB)	26.40	42.36	50.42	65.76
Enc-base (GMDB)	**28.25**	**44.88**	**52.00**	**66.18**

Table 2. Few-shot meta baseline and feature-level fusion on GMDB (v1.0.3) retrieval task.

Method	Top-1	Top-5	Top-10	Top-30
Frequent Set				
GMDB-fs	48.06	68.13	75.89	85.67
(feature-level fusion)				
+Enc-healthy	47.55	**68.47**	**77.23**	**88.69**
+Enc-all (GMDB)	**47.55**	67.62	74.20	84.65
+Enc-base (GMDB)	47.22	67.96	74.71	84.82
Rare Set				
GMDB-fs	30.21	48.19	56.39	71.07
(feature-level fusion)				
+Enc-healthy	32.89	**50.65**	**57.89**	**71.39**
+Enc-all (GMDB)	30.88	48.29	56.57	70.54
+Enc-base (GMDB)	**33.08**	48.37	56.65	70.72

Few-shot meta-training (GMDB-fs) improves the top-1 frequent test accuracy of the best GMDB-trained baseline models, Enc-all and Enc-base by 7.59%, and 5.56%, respectively. This improvement is not limited to top-1 retrieval, it is also retained in different neighbor retrieval. We initialized GMDB-fs models using healthy encoding.

Table 3. Comparison of n categories with 4 tasks per episode and 10 categories with n-shot and n-query.

	Frequent				Rare			
	Top-1	Top-5	Top-10	Top-30	Top-1	Top-5	Top-10	Top-30
n-categories 3-shot / 2-query								
5	**49.24**	66.44	75.04	84.49	27.70	54.41	54.44	69.37
10	48.06	**68.13**	**75.89**	85.67	30.21	48.19	56.39	71.07
15	47.89	67.96	75.72	**86.51**	**31.63**	**49.35**	**58.17**	**72.95**
20	48.06	67.62	74.37	84.65	27.76	47.04	55.29	69.33
n-shot/n-query, 10-categories								
1/4	44.35	65.94	73.19	84.65	29.37	46.99	56.26	69.47
2/3	44.35	67.12	74.87	**86.34**	**31.77**	**49.26**	**57.66**	71.04
3/2	47.05	**69.14**	**76.05**	**86.34**	30.15	48.38	57.17	**71.34**
4/1	**48.23**	68.13	75.21	84.65	27.76	45.79	55.58	68.72

In both frequent and rare sets, feature-level fusion with the healthy encoder performed the best in nearly all retrieval tasks. In top-1 rare retrieval, fusion with Enc-base gives the best accuracy, 33.08%. Even though Enc-all and Enc-base perform better than Enc-healthy, their performance on feature fusion is limited.

4. Discussion

We observed differences in the model's behavior when evaluating the few-shot meta-based training with different sets of configurations (Table 3). These variables affect the difficulty of few-shot tasks and must be examined in depth. One is the number of ways

(categories) to define possible classes in a support set. The best retrieval performance is in the 10 and 15 categories. We consider this behavior related to the complexity of classification tasks in each episode. We picked the 10-way to evaluate other parameters that affect the performance of episodic training. These are the number of images in each class in the support set (k-shot) and the number of query images. Table 3 (bottom) presents evaluation performance using the different number of shots and queries. A higher number of shots improves frequent set retrieval performance; however, the 4-shot setting performs worse in rare set retrieval. The minimum number of shots seems descriptive enough according to the n-way task learned. We could not increase n-shot and queries as the minimum number of samples per class in the training set was 5. The best overall performance is in 2-shot or 3-shot settings.

5. Conclusion

In this study, we trained a state-of-the-art face recognition model on standard face recognition databases. We transferred learned representations to our low-resource target data domain for facial phenotype recognition for genetic disorders. We addressed the issue of data scarcity and imbalanced data using a few-shot meta-based learning approach. This improved genetic disorder recognition of unseen genetic conditions compared to the recently published GestaltMatcher DCNN; however, our study has certain limitations. We need more samples of rare diseases to ensure a fine-grained analysis and a user study of how AI models and clinicians' decisions deviate. In future work, using generative models on either image or feature level, synthesized samples can also be added to few-shot training and reduce the effect of uneven class distribution.

References

[1] Baird PA, Anderson TW, Newcombe HB, Lowry RB. Genetic disorders in children and young adults: a population study. American journal of human genetics. 1988; 42(5): 677-693.
[2] Hart TC, Hart PS. Genetic studies of craniofacial anomalies: clinical implications and applications. Orthodontics & craniofacial research. 2009 February; 12(3): 212-220.
[3] Ferry , Steinberg J, Webber C, FitzPatrick DR, Ponting CP, Zisserman A, et al. Diagnostically relevant facial gestalt information from ordinary photos. eLife. 2014 June.
[4] Loos HS, Wieczorek D, Würtz RP, Malsburg Cvd, Horsthemke B. Computer-based recognition of dysmorphic faces. European Journal of Human Genetics. 2003; 11(8): 555-560.
[5] Shukla P, et al. A deep learning framework for recognizing developmental disorders. In 2017 IEEE Winter Conference on Applications of Computer Vision (WACV); 2017: IEEE. p. 705-714.
[6] Gurovich Y, Hanani Y, Bar O, Guy N, Fleischer N, Gelbman D, et al. Identifying facial phenotypes of genetic disorders using deep learning. Nature medicine. 2019; 25(1): 60-64.
[7] Hsieh TC, Bar-Haim A, Moosa S, Ehmke N, Gripp KW, Pantel JT, et al. GestaltMatcher facilitates rare disease matching using facial phenotype descriptors. Nature genetics. 2022 February; 54(3): 349-357.
[8] Deng J, Guo J, Xue N, Zafeiriou S. Arcface: Additive angular margin loss for deep face recognition. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2019. p. 4690-4699.
[9] Snell J, Swersky K, Richard Z. Prototypical networks for few-shot learning. In Advances in neural information processing systems; 2017: Curran Associates, Inc.
[10] Gidaris S, Komodakis N. Dynamic few-shot visual learning without forgetting. In Proceedings of the IEEE conference on computer vision and pattern recognition; 2018. p. 4367-4375.
[11] Yinbo C, Liu Z, Xu H, Darrell T, Wang X. Meta-baseline: Exploring simple meta-learning for few-shot learning. In Proceedings of the IEEE/CVF International Conference on Computer Vision; 2021. p. 9062.