Facial Skin Classification Using Convolutional Neural Networks

Jhan S. Alarifi(✉), Manu Goyal, Adrian K. Davison, Darren Dancey, Rabia Khan, and Moi Hoon Yap

Manchester Metropolitan University, Manchester M1 5GD, UK
Jhan.s.alarifi@mmu.ac.uk

Abstract. Facial skin assessment is crucial for a number of fields including the make-up industry, dermatology and plastic surgery. This paper addresses skin classification techniques which use conventional machine learning and state-of-the-art Convolutional Neural Networks to classify three types of facial skin patches, namely normal, spots and wrinkles. This study aims to accomplish the pivotal work on the basis of these three classes to provide the collective facial skin quality score. In this work, we collected high quality face images of people from different ethnicities to create a derma dataset. Then, we outlined the skin patches of 100 × 100 resolution in the three pre-decided classes. With extensive parameter tuning, we ran a number of computer vision experiments using both traditional machine learning and deep learning techniques for this 3-class classification. Despite the limited dataset, GoogLeNet outperforms the Support Vector Machine approach with *Accuracy* of 0.899, *F-Measure* of 0.852 and *Matthews Correlation Coefficient* of 0.779. The result shows the potential use of deep learning for non-clinical skin images classification, which will be more promising with a larger dataset.

Keywords: Facial skin · CNNs · Classification · Skin quality assessment

1 Introduction

Data on skin quality properties are often assembled and evaluated by a well-trained expert who allocates noticeable skin samples, both live or from photographs, to a recognised quality grade on a predefined grading scale. However, a machine vision approach to assess skin quality properties is useful in providing an objective analysis [1,2]. This can avoid problems with repeatability and reproducibility, since a professional's experience and knowledge is subjective and can differ amongst graders. This can also potentially result in reduced cost and more effective analysis while providing a consistent assessment of skin quality [3]. There is great importance in providing objectivity to the dermatologist's visual evaluation of skin in order to efficiently develop effective pharmaceutical treatments. Recently, several skin assessment methods have been established; for instance, analysis of the skin appearance around pores on the face [4], evaluation of facial wrinkle improvements over time [5], measuring facial wrinkles using

F. Karray et al. (Eds.): ICIAR 2017, LNCS 10317, pp. 479–485, 2017.
DOI: 10.1007/978-3-319-59876-5_53

quantification methods and automatic detection [6]. Most of these assessments were subjective and revolved around a clinical perspective and a professional's opinion rather than an objective assessment. Further research is required to understand the definition of skin quality based on human perception. In this work, we have experimented with several conventional machine learning (CML) methods and Convolutional Neural Networks (CNNs) with different parameters and settings to classify spots, wrinkles and normal skin patches. This was followed by a comparison of Support Vector Machine (SVM) [7] and GoogLeNet [8] performances.

The rest of the paper is organised as follows: The related work on classification of skin is given in Sect. 2; Sect. 3 elaborates on the dataset and experimental settings; Sect. 4 evaluates the efficiency of the different classifiers that are the best fit for the proposed purpose, for instance, *Sensitivity* and *F-Measure*; in the final section, the conclusion presents the prospects for future work and the limitations of this work.

2 Related Work

Standard machine learning methods have been widely used in several pattern recognition tasks. They have also been used for the detection of skin conditions such as acne [9]. These traditional machine learning methods performed well in many classification tasks. However, they do come with some consequences. For example, ANN (Artificial Neural Network) can be affected by the number of hidden layers, hidden nodes and learning rates. Another disadvantage is that the network has to be extensively trained in order to achieve optimal performance, which is why SVM was chosen for this experiment as a more suitable option. SVM has been used commonly over the last decade [10]. A categorisation of skin texture in early melanoma detection method was implemented using SVM and for skin colour categorisation [11,12]. However, is the Convolutional Neural Networks (CNNs), which is a deep learning framework, has outperformed other method in image classification domain [10].

Recently, with the rapid growth of deep learning algorithms, they become most effective in classification tasks such as in facial recognition and face tracking. The purpose is to understand hierarchical representations of data by using a deep architecture model [13]. Krizhevsky et al. [14] used a deep convolutional neural network to classify high-resolution images in the ImageNet LSVRC-2010. Therefore, including deep learning in the process would provide better performance and more reliable results for the desired output. The network was trained with a total of 1.2 million images and 1000 different classes with error rates of 39.7% for top 1 and 18.9% for top 5. This illustrates the advantage of using this approach. On the other hand, the data used in that approach do not relate to skin attributes. Andre et al. [15] applied successful deep learning approach of skin cancer to dermatologist level by comparing the network performance against 21 dermatologists. Nevertheless, the research focused on clinical use. Therefore, this work will observe the performance of using CNNs in classification of non-clinical skin features such as spots and wrinkles.

3 Methodology

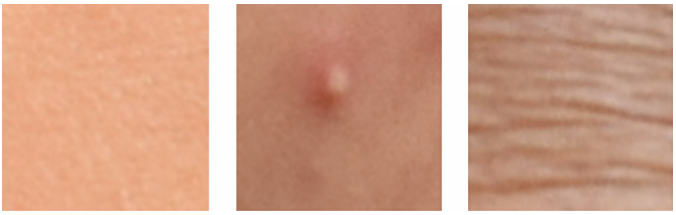
This section will describe in depth the appropriate datasets available, the two sets of experiments and their settings.

3.1 Dataset

Currently, there are limited datasets available for the analysis of facial skin conditions. An available dataset called DermNet consists of a total of 23000 images of various skin diseases. However, this dataset has two limitations. One limitation was that the data collection was not under a controlled environment, which has caused inconsistencies in the images and affected their integrity as well as their accuracy. Another limitation was that the images were not only of facial skin conditions, but also of different diseased body parts, which are unsuitable for this experiment focusing on the classification of common facial skin conditions. To address these limitations, we proposed an ongoing collection of consistent, high-quality images of faces from a wide demographic and from participants who engage in different social habits. These habits can include, but are not limited to, smoking and alcohol consumption. The dataset currently consists of 164 images of participants with a mean age of 48.43 (standard deviation (SD): 21.44, ages between 18 and 92). There are 25 different self-reported ethnicities in the dataset including African, Arabic, Chinese and Malaysian. The ethnic group with most participants is White British with 119 images. The main reported gender was female with a total number of 107 participants; there were also 56 male participants next and 1 transgender participant. To understand how certain habits can affect a person's facial skin properties, participants were asked to complete a questionnaire asking if they consumed alcohol or smoked. Overall, 68 participants never drank alcohol, 88 currently drink and 8 used to drink but had stopped. As for smoking, 85 people never smoked, 21 currently smoke tobacco in some form, 1 smokes electronic cigarettes only, 6 had partaken in smoking a few times in their lives and 51 used to smoke but stopped. The images were taken with a Nikon D5300 at a resolution of 4496×3000 to ensure that as much detail as possible on participants' faces was captured.

Firstly, five expressionless images of each participant were captured at different angles to allow for a full view of the face and its profiles. Next, participants were asked to pose with six different facial expressions which were based on Ekmans [16] universal facial expressions: happiness, sadness, surprise, disgust, anger and fear. The replication of these expressions allows for the dataset to include within it some variation in the way each participant's facial skin changes due to natural expressions. Being able to differentiate between actual wrinkles and ridges caused by expression lines would be extremely useful when analysing facial conditions in the future, as would the ability to distinguish between changes caused by natural expressions and deformities caused by other reasons like aging and social habits. The dataset is an on-going project of data collection. Therefore, in the near future, the dataset is likely to increase in size.

Skin patches were also collected. These were of size 100 × 100 and consisted of three categories: normal skin, skin with spots and skin with wrinkles, as illustrated in (Fig. 1). The spotted skin class has different stages of spots, inflamed and non-inflamed. The wrinkled skin class has two different types of wrinkles: deep and fine wrinkles, which were taken from different parts of the face. The total number of patches is 325.



(a) Normal (b) Spot (c) Wrinkle

Fig. 1. Sample skin patch from each three classes.

3.2 Traditional Machine Learning

In this section, we used traditional supervised machine learning for the classification of three classes (Normal, Spot, and Wrinkle). Since these three classes of skin have major textural differences amongst them, we investigated popular feature extraction techniques including texture descriptors such as Local Binary Patterns (LBP) [17], and Histogram of Oriented Gradients (HOG) [14]. We did not include mutltifractal as texture descriptor due to it is better in representing face features than skin region [18]. In addition, we also used color descriptors such as Normalized RGB, HSV, and L*u*v features. After the feature extraction from images, we used the machine learning classifier Sequential Minimal Optimization (SMO) to train Support Vector Machine (SVM) for classification task.

3.3 Deep Learning

The Caffe framework [19] was chosen to implement the state-of-the-art CNNs architecture of GoogLeNet. The intention was to provide improvements on the existing model AlexNet when it comes to classifying ImageNet [8]. This model contains 22 layers, compared to AlexNet and CaffeNets [20]. To investigate the best optimisation algorithm for skin patches classification, we tested a number of solvers such as Stochastic Gradient Descent (SGD), Nesterovs Accelerated Gradient (NAG) and Adaptive Gradient (AdaGrad). SGD is one of the most commonly used approaches for large-scale machine learning tasks [21]. AdaGrad presented strong experimental performance on real-world complications, which were tested under different parameters as follows: Each optimizer was tested on the default setting using 30 epochs and 0.01 learning rate. On the second tested

set, the number of epochs was increased to 60 and the learning rate was kept the same. For the last tested set, the learning rate was decreased to 0.001 and the number of epochs was kept at 60 [22]. Since the data starts to converge, there is no need to increase the number of epochs.

4 Results and Discussion

In this section, we present the results for various classification experiments on our face dataset of 164 images. These high-resolution images were manually split into the three pre-defined classes of skin patches with 100 × 100 resolution. For this 3-class classification, we divided the dataset of skin patches into 70% for the training set and 30% for the testing set. We adopted the 10-fold cross validation technique to create 10 test cases with a total of 325 images of skin patches. We then divided the number of images equally from the three categories of skin patches. Thus there were 228 skin patches for the training dataset, 97 for testing set. As for evaluation. We chose a number of popular performance measures such as *Sensitivity*, *False Negative Rate (FNR)*, *F-Measure*. *Recall*, *Precision*, *Matthews Correlation Coefficient (MCC)*, and *Accuracy*. We investigated both traditional machine learning and extensive deep learning techniques to carry out the classification experiments.

Table 1. SVM resuts

Method	*Sensitivity*	*F-Measure*	*Recall*	*Precision*	*MCC*	*Accuracy*
LBP	**0.742**	**0.741**	0.741	**0.740**	**0.597**	**0.815**
LBP and HOG	0.736	0.738	**0.742**	0.742	0.591	0.811
LBP, HOG and Colour Descriptor	0.733	0.735	0.740	0.740	0.586	0.808

Tables 1 and 2 show the classification results achieved with the help of traditional machine learning techniques and deep learning techniques respectively. It is clearly illustrated that with proper parameter tuning, the deep learning techniques outperformed the traditional machine learning ones within the dataset used. Though deep learning techniques usually require a large dataset to train models for classification, this limited dataset was still able to achieve an accuracy rate of 85% and *Sensitivity*, *Recall*, *Precision* and *MCC* rates of 0.854, 0.856, 0.856, and 0.779 respectively. This is promising since traditional machine learning techniques are only able to get the best accuracy of approximately 74%. NAG is a first order method and has a distinctive mechanism compared to gradient descent in certain conditions in terms of convergence rate [21]. This predicts the gradient for the next epoch and updates the learning rate for the existing iteration based on the predicted gradient. Therefore, if the gradient is increased for the next set, the learning rate for the present iteration would be higher.

Table 2. GoogLeNet results.

Solver	*Epochs*	*Learning rate*	*Sensitivity*	*F-Measure*	*Recall*	*Precision*	*MCC*	*Accuracy*
SGD	30	0.01	0.666	0.661	0.666	0.666	0.472	0.754
	60	0.01	0.833	0.835	0.835	0.835	0.745	0.884
	60	0.001	0.677	0.670	0.671	0.671	0.487	0.761
NAG	30	0.01	0.646	0.639	0.645	0.645	0.439	0.738
	60	**0.01**	**0.854**	**0.852**	**0.856**	**0.856**	**0.779**	**0.899**
	60	0.01	0.729	0.727	0.731	0.732	0.579	0.856
AdaGrad	30	0.01	0.521	0.425	0.375	0.375	0.192	0.624
	60	0.01	0.646	0.650	0.667	0.667	0.449	0.739
	60	0.001	0.708	0.703	0.707	0.707	0.545	0.790

Conversely, if the gradient is low, it would slow down the learning rate. In this experiment, the solver received the highest accuracy with 60 epochs and default learning rate.

5 Conclusion

We presented a dataset that is suitable for facial skin analysis. Our experiments showed the potential for using CNNs in classifying skin attributes. Thus far, GoogLeNet using NAG outperforms the other optimisers used in the experiments. Although the data collection was under a controlled environment and had high-resolution images, it is limited to three categories. Therefore, an expansion of the data is needed.

To improve the classification accuracy for non-clinical skin images, future research involves conducting experiment to understand human perception in classifying skin types and collect more data. We are also interested in comparing the performance of the experts and non-experts [23], in this case, the differences between dermatologists to non-dermatologists.

References

1. Ng, C.-C., Yap, M.H., Costen, N., Li, B.: Automatic wrinkle detection using hybrid Hessian filter. In: Cremers, D., Reid, I., Saito, H., Yang, M.-H. (eds.) ACCV 2014. LNCS, vol. 9005, pp. 609–622. Springer, Cham (2015). doi:10.1007/978-3-319-16811-1_40
2. Ng, C.-C., Yap, M.H., Costen, N., Li, B.: Wrinkle detection using Hessian line tracking. IEEE Access **3**, 1079–1088 (2015)
3. Prats-Montalbán, J.M., Ferrer, A., Bro, R., Hancewicz, T.: Prediction of skin quality properties by different multivariate image analysis methodologies. Chemometr. Intell. Lab. Syst. **96**(1), 6–13 (2009)
4. Mizukoshi, K., Takahashi, K.: Analysis of the skin surface and inner structure around pores on the face. Skin Res. Technol. **20**(1), 23–29 (2014)

5. Luebberding, S., Krueger, N., Kerscher, M.: Comparison of validated assessment scales and 3D digital fringe projection method to assess lifetime development of wrinkles in men. Skin Res. Technol. **20**(1), 30–36 (2014)
6. Cula, G.O., Bargo, P.R., Nkengne, A., Kollias, N.: Assessing facial wrinkles: automatic detection and quantification. Skin Res. Technol. **19**(1), e243–e251 (2013)
7. Wang, L.: Support Vector Machines: Theory and Applications, vol. 177. Springer Science & Business Media, Heidelberg (2005)
8. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–9 (2015)
9. Liao, H.: A deep learning approach to universal skin disease classification
10. Schmidhuber, J.: Deep learning in neural networks: an overview. Neural Netw. **61**, 85–117 (2015)
11. Yuan, X., Yang, Z., Zouridakis, G., Mullani, N.: SVM-based texture classification and application to early melanoma detection. In: 28th Annual International Conference of the IEEE, Engineering in Medicine and Biology Society, EMBS 2006, pp. 4775–4778. IEEE (2006)
12. Khan, R., Hanbury, A., Stöttinger, J., Bais, A.: Color based skin classification. Pattern Recogn. Lett. **33**(2), 157–163 (2012)
13. Wang, L., Sng, D.: Deep learning algorithms with applications to video analytics for a smart city: a survey. arxiv preprint (2015). arXiv:1512.03131
14. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems, pp. 1097–1105 (2012)
15. Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., Thrun, S.: Dermatologist-level classification of skin cancer with deep neural networks. Nature **542**, 115–118 (2017)
16. Ekman, P.: Facial expressions. Handb. Cogn. Emot. **16**, 301–320 (1999)
17. Guo, Z., Zhang, L., Zhang, D.: A completed modeling of local binary pattern operator for texture classification. IEEE Trans. Image Process. **19**(6), 1657–1663 (2010)
18. Yap, M.H., Ugail,H., Zwiggelaar, R., Rajoub, B., Doherty, V., Appleyard, S., Hurdy, G.: A short review of methods for face detection and multifractal analysis. In: International Conference on CyberWorlds, CW 2009, pp. 231–236. IEEE (2009)
19. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: convolutional architecture for fast feature embedding. In: Proceedings of the 22nd ACM International Conference on Multimedia, pp. 675–678. ACM (2014)
20. Liu, D., Wang, Y.: Monza: image classification of vehicle make and model using convolutional neural networks and transfer learning
21. Singh, B., De, S., Zhang, Y., Goldstein, T., Taylor, G.: Layer-specific adaptive learning rates for deep networks. In: 2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA), pp. 364–368. IEEE (2015)
22. Duchi, J., Hazan, E., Singer, Y.: Adaptive subgradient methods for online learning and stochastic optimization. J. Mach. Learn. Res. **12**, 2121–2159 (2011)
23. Yap, M.H., Edirisinghe, E., Bez, H.: Processed images in human perception: a case study in ultrasound breast imaging. Eur. J. Radiol. **73**(3), 682–687 (2010)