A Benchmark for Automatic Visual Classification of Clinical Skin Disease Images

Xiaoxiao Sun, Jufeng Yang(✉), Ming Sun, and Kai Wang

College of Computer and Control Engineering, Nankai University, Tianjin, China
yangjufeng@nankai.edu.cn

Abstract. Skin disease is one of the most common human illnesses. It pervades all cultures, occurs at all ages, and affects between 30 % and 70 % of individuals, with even higher rates in at-risk. However, diagnosis of skin diseases by observing is a very difficult job for both doctors and patients, where an intelligent system can be helpful. In this paper, we mainly introduce a benchmark dataset for clinical skin diseases to address this problem. To the best of our knowledge, this dataset is currently the largest for visual recognition of skin diseases. It contains 6,584 images from 198 classes, varying according to scale, color, shape and structure. We hope that this benchmark dataset will encourage further research on visual skin disease classification. Moreover, the recent successes of many computer vision related tasks are due to the adoption of Convolutional Neural Networks(CNNs), we also perform extensive analyses on this dataset using the state of the art methods including CNNs.

Keywords: Skin disease image · Computer aided diagnosis · Image classification · CNNs · Hand-crafted features

1 Introduction

Skin disease is one of the most common illnesses in human daily life. It pervades all cultures, occurs at all ages, and affects between 30 % and 70 % of individuals [1]. There are tens of millions of people affected by it every day. Skin disease is twofold, i.e. skin infection and skin neoplasm, in which thousands of skin conditions have been described [2]. Skin disease has a major adverse impact on quality of life and many are associated with significant psychosocial mobility. However, only a small proportion of people can recognize these diseases without access to a field guide. Moreover, there are many over-the-counter (OTC) drugs to treat the frequently-occurring skin diseases in daily life. In this case, correctly recognizing the skin diseases becomes very important for people who need to make a choice about these medicines. If people want to make a preliminary self diagnosis, it is undisputed that a visual recognition system will be useful for assisting them even if it is not perfect. For example, if an accurate skin disease classifier is developed, a user can submit a photo of recently skin condition to query a diagnosis. Surprisingly, there exists few research using computer vision techniques to recognize many common skin diseases based on ordinary photographical images.

B. Leibe et al. (Eds.): ECCV 2016, Part VI, LNCS 9910, pp. 206–222, 2016.
DOI: 10.1007/978-3-319-46466-4_13

Fig. 1. Examples of dermoscopic and clinical images. (a) Dermoscopic images are acquired through a digital dermatoscope, which have relatively low levels of noise and consistent background illumination. (b) Clinical images are collected via various sources, most of which are captured with digital cameras and cell phones

Despite there are some related applications, the problem of recognizing skin diseases has not been fully solved by the computer vision community. In contrast to object or scene classification, skin disease image has no distinctive spatial layout, as we can label a bird with its body and head or an outdoor scene with sky region and house. For example, it's difficult for us to find an accurate description of scattered red eczema. Besides, there are many challenges, including low contrast between lesion and surrounding skin, irregular and fuzzy borders, fragmentation or variegated coloring inside the lesion, etc., which make it hard to recognize skin diseases.

Most previous works on recognition of skin disease are restricted to dermoscopic images [3,4], which are acquired through a digital dermatoscope. A dermatoscope is a special device for dermatologists to use to look at skin lesions that acts as a filter and magnifier [5]. As a result, dermoscopic images have low level of noise and are always with unique lighting. We show some examples of dermoscopic images in Fig. 1(a). On the other hand, clinical skin disease images are collected via a variety of sources, most of which are acquired using digital cameras and cell phones. Examples are shown in Fig. 1(b). We have found some work based on clinical disease images [5,6]. However, all these work are built on small datasets which only contain very few species and are not publicly available. The absence of benchmark datasets is a barrier to a more dynamic development of this research area. As a consequence, in this paper, we introduce a new, publicly available dataset for real-world skin disease images recognition. This dataset contains 6,584 images of 198 fine-grained skin disease categories.

As is well known, image classification is one of the most fundamental problems of computer vision, and has been studied for many years. Large-scale annotated image datasets have been instrumental for driving progress in object recognition over the last decade. These datasets contain a wide variety of

basic-level classes, such as different kinds of animals and inanimate objects. Significant progress has been made in the past few years in object classification as researchers compete on the ImageNet Large Scale Visual Recognition Challenge (ILSVRC).

Compared to generic object classification, fine-grained visual categorization [7–11] aims to classify categories which belong to the same basic-level class. In recent years, fine-grained recognition has been demonstrated in many domains with corresponding datasets, including birds [12,13], flowers [14,15], leaves [16], dogs [17,18], and cars [19]. A variety of methods have been developed for classifying fine-grained categories [7,11,20–23].

Skin disease image classification is naturally considered belonging to the problem of fine-grained visual object classification. However, in contrast to scene classification or object classification, it has own characteristics different from the existing fine-grained classification work, because it's a difficult problem that push the limits of the visual abilities for both human and computers. Clinically, the diagnosis of any particular skin condition is made by gathering pertinent information regarding the presenting skin lesion(s), including the location, symptoms, duration, arrangement (solitary, generalized, annular, linear), morphology (macules, papules, vesicles), and color (red, blue, brown, black, white, yellow) [24]. In addition, the diagnosis of many conditions often requires more complicated information.

In order to validate the usefulness of our proposed dataset and inspire the computer vision community to carry out more meaningful research in this field, we perform a lot of basic experiments employing both hand-crafted features and deep features to establish a baseline performance on the dataset. On the other hand, recently deep learning has enabled robust and accurate feature learning, which in turn produces the state-of-the-art performance on many computer vision related tasks. In this work, we want to find out whether or not applying CNNs to skin disease classification provides advantages over hand-crafted features.

Our contributions are summarized as follows. First, we collect a novel and large scale benchmark dataset for skin disease image recognition. Second, we evaluate the performance of skin disease classification using CNNs as well as hand-engineered features. Extensive experimental results show that using the existing CNN model does not outperform manually crafted visual features. On the other hand, we hope this can promote future research on skin disease classification with deep learning.

2 Related Work

Our work is closely related to image classification on both dermoscopic and clinical images, and convolutional neural networks.

2.1 Dermoscopic Images Recognition

Present works on skin disease image classification are twofold, that is, dermoscopic and clinical image recognition. First, we introduce the representative works on dermoscopic images.

Dermoscopic images have been mostly used in computer aided diagnosis, which is a technique of visualizing lesions by directing light onto the skin. Because dermoscopic images have bright illumination conditions, it is clear enough for recognition. Besides, the viewpoint is basically invariable and background clutter is very limited. All these characteristics make the processing of dermoscopic images easier, which further result that the computer vision studies based on dermoscopic images are much more than work based on clinical images.

Some work have focused on developing different components of dermoscopic image recognition, including segmentation [25], detection [26] and classification [3,27], etc. Gonzalez-Castro *et al.* [3] introduce a color texture descriptor and apply it to classify images of nevi into benign lesions and melanoma. Celebi *et al.* [27] present a methodological approach for the classification of dermoscopy images. The approach involves border detection, feature extraction, and SVM classification with model selection. Kasmi and Mokrani [28] introduce an algorithm that extracts the characteristics of ABCD (asymmetry, border irregularity, colour and dermoscopic structure) attributes to build a binary classifier, again distinguishing melanoma from benign nevus.

The popular datasets of dermoscopic images used in recent works are shown in Table 1. There is no doubt that these studies have developed the diagnosis of skin diseases. However, their applications are limited due to the specialized medical equipments and requirement for expert knowledge. Different from the mentioned datasets, in this paper, we build a large scale clinical image dataset to encourage further research which could be applied in real life scenes.

2.2 Clinical Image Recognition

Some efforts have been made to classify clinical skin disease images [35–37]. Concretely, Glaister *et al.* [5] propose a segmentation algorithm based on texture distinctiveness (TD) to locate skin lesions in photographs. They introduce a joint statistical TD metric and a texture-based region classification algorithm,

Table 1. Statistics of recent datasets of dermoscopic images. Also, the representative work employing these datasets are listed here.

Dataset	[29]	[27]	[30]	[31]	[32]	[33]	[34]	[25]	[28]
Classes#	2	2	6	2	3	2	2	2	2
Images#	527	596	320	1097	945	241	200	208	200
Year	2000	2007	2011	2012	2013	2015	2015	2015	2016
Available?	Y	N	Y	N	Y	N	Y	N	N

which captures the dissimilarity between learned representative texture distributions. Alcón *et al.* [38] describe an automatic system for inspection of pigmented skin lesions and discriminating between malignant an benign lesions. The system includes a dedicated image processing system for feature extraction and classification, and patient-related data decision support machinery for calculating a personal risk factor. It has been shown that their algorithm is capable of recreating controlled lighting conditions and correcting for uneven illumination.

Moreover, Razighi *et al.* [6,39,40] heavily rely on human-in-the-loop and high level knowledge in their work. They use human provided information with a random forest or bayesian framework. The aforementioned interaction information comes from questions designed in advance. For example, the answer to a binary question like: Is the object red? can be regarded as the presence of tag *Red*, that can be used as a visual feature to improve the final classification result. They include 10 questions and 37 possible binary answers/tags in their system.

Typically, the previous works focusing on clinical skin disease images are commonly built on a small size datasets. To the best of our knowledge, the largest dataset contains 2309 images from only 44 different diseases, and it is not publicly available to the community.

2.3 Convolutional Neural Networks

In recent years, Convolutional Neural Networks (CNNs) have achieved great empirical successes in many computer vision tasks, such as image classification [41], object detection [42], scene recognition [43], and fine-grained classification [23,44,45]. It is now possible to train a very deep network [46] on large collections of images with the help of the increasing computational power of GPU.

Skin diseases have the similar characteristics with objects in fine-grained classification, that is, lesion areas in skin disease images show large intra-class variation and small inter-class variation. Therefore, CNNs are also supposed to make sense in skin disease recognition. On the other hand, skin disease images are different from conventional fine-grained object images in some degrees. For example, some current works in fine-grained classification employ bounding box of objects of interest to help recognition, while it's more difficult label bounding box in skin disease images, or to distinguish lesions from background. Furthermore, objects always have specific shapes and parts, resulting a massive of part based methods to train fine-grained part models in CNNs. However, choosing parts of lesion is almost impossible when skin disease images are applied.

3 Our Dataset

Several datasets have been used for skin disease studies [6,37]. Concretely, Razeghi *et al.* [37] collect two subsets in their work, which contain 90 and 706 images from 3 and 7 different skin diseases, respectively. In another work of the same team [6], they acquire a new dataset containing 2,309 visual similar images

Table 2. Statistics of the existing clinical skin disease datasets. In [6,37], the authors add a question and answer bank into their datasets, which is used to provide human-computer interaction in the systems. Note that none of current datasets are publicly available. For comparison, we also show information of our work in the last two columns, which expand both the dataset size and the number of categories.

Dataset	[37]-1^{st}	[37]-2^{nd}	[6]	SD-128	SD-198
Classes#	3	7	44	128	198
Images#	90	706	2309	5619	6584
Year	2012	2012	2013	2016	2016
Available?	N	N	N	Y	Y

of skin conditions from 44 different diseases. The authors argue that the lesions are manually segmented using a bounding box in their dataset, and the dataset has a question and answer bank for help classification. Unfortunately, both of the mentioned datasets are not publicly available.

In this work, we present a new clinical skin disease dataset, namely SD-198. To the best of our knowledge, it is the largest available skin disease database, whether clinical or dermoscopic images are mentioned. The statistics of the existing skin disease datasets are shown in Table 2.

Our SD-198 dataset contains 198 different diseases from different types of eczema, acne and various cancerous conditions. There are 6,584 images in total. We also choose the classes with more than 20 image samples as a subset, namely SD-128. In general, overall classification can be improved when less categories and more samples are applied, which is verified in our experiments. Examples of images in our dataset can be found in Fig. 2.

3.1 Image Collection and Annotation

Collection. The images are downloaded from the DermQuest[1], which is an online medical resource for dermatologists and healthcare professionals with an interest in dermatology. It contains an extensive clinical images shared by the wide dermatology community. The images are submitted by patients or dermatologists.

The website contains 729 species of skin lesions in total, which include all kinds of conditions that affect the integumentary system, i.e., the organ system that encloses the body and includes skin, hair, nails, and related muscle and glands [47]. We execute a statistical analysis of these skin lesion categories, and remove the species that rarely appear in real life or that have less than 10 samples.

We initially have collected more than 10,000 clinical skin disease images. In order to keep balance of categories, we remove some samples from the subsets whose images are sufficient so as to each category has 60 images at most. Then,

[1] https://www.dermquest.com.

Fig. 2. Here we show some examples of our SD-198 dataset, each of which is selected from different classes. In another word, none of the images listed here have the same class label. However, it's difficult to distinguish these skin disease images, because some of them have the extremely same color and shape. For example, the five images in the first column belong to different categories, while finding the differences among these images are challenging. (Color figure online)

we further drop the duplicate images and low-quality images. Finally, we get a dataset containing 6,584 images from 198 different categories.

Annotation. The ground truth annotations of the images in our collected dataset are obtained from DermQuest, since each image has been recognized by experts and labeled with the name of its class. Because the clinical case notes and diagnosis quizzes on the website are reviewed by an international editorial board comprised of renowned dermatologists, the labels obtained for our dataset are considered reliable. Despite that, in order to ensure the label quality of our dataset, we have invited two professionals to review our dataset.

3.2 Properties of Dataset

Not only our dataset is larger than previous datasets, but also has superior performance. We will introduce the properties of the proposed dataset in this section.

Scale. This paper aims to provide a large-scale clinical skin disease benchmark dataset. To the best of our knowledge, its size is about 3 to 10 times as the reported scale of the previous datasets. It contains 198 categories which have covered all of the common skin diseases. We hope the dataset with 6,584 well-labeled clinical skin disease images can promote the vision research in this area.

Diversity. All images are from the real scene with variance in color, exposure, illumination and level of details. That is to say the images may be taken by any configuration of equipments or in a variety of environments. Therefore, hopefully the future works based on our benchmark dataset will be easier to be applied into practices. The mentioned diversities mainly include:

(1) Species Diversity: Skin lesion images in our dataset contain: eczema, psoriasis, acne vulgaris, pruritus, alopecia areata, decubitus ulcer, urticaria, scabies, impetigo, abscess, bacterial skin diseases, viral warts, molluscum, melanoma and non-melanoma skin cancer, which have covered most of the common skin diseases. Figure 3 shows the statistics of the number of images in each class.

We also show some images in Fig. 4. For example, in Fig. 4(a), the first row is angioma, and the second row contains four kinds of diseases. In the third row, images of acne vulgaris and guttate psoriasis are in green and yellow boxes, respectively. Figure 4(b) also contains three kinds of diseases. Due to space limitation, we do not show all classes. However, one can already find the species diversity of clinical skin disease images in these figures.

(2) Appearance Diversity: In real life, clinicians and dermatologists determine whether a lesion is a melanoma by a certain criteria, that is ABCD criteria (asymmetry, border irregularity, colour and diameter or differential structures). The criteria is proposed by Friedman *et al.* [48], which has been widely adopted through the previous works, especially in dermoscopic image recognition.

Compared to dermoscopic images, there are different meanings of ABCD in clinical skin disease images. We summarize ABCD's conventional meanings and refine them to apply to clinical images in our dataset. In Fig. 4(b), we show some examples of clinical images based on ABCD criteria. The images in the same row represent the A-asymmetry, B-border irregularity, and C-multiple colors,

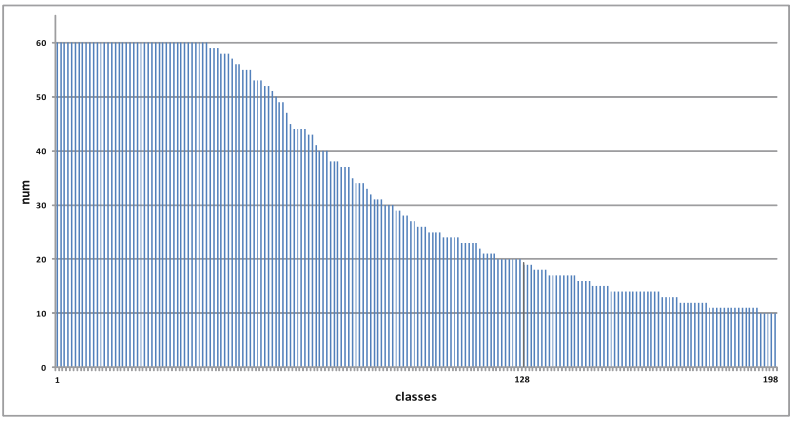

Fig. 3. Statistics of the numbers of images for each class in our SD-128 and SD-198 datasets. Note that each category of SD-128 contains more than 20 samples, while SD-198 has some categories whose samples are between 10 and 20.

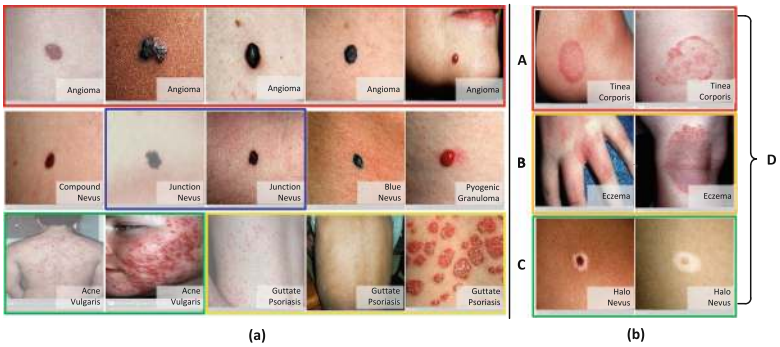

Fig. 4. Species diversity and appearance diversity of our proposed dataset. If I tell you that the images in the first row of (a) belong to the same class, do you think the images in the next row are from the same class? The answer is no. Moreover, the third row of (a) show that different shooting distance and illumination have a big influence on the appearance of skin diseases. In (b) we show some examples with the ABCD criteria. Note that these mentioned diversities, as well as attribute diversity, contribute to making automatic recognition of clinical skin disease image a challenging work. (Color figure online)

respectively. The D-diameter is difficult to be judged by images, but we can see from Fig. 4(b) that it varies greatly among different diseases.

Skin diseases in our dataset show that they have different appearances from an ABCD perspective, which includes arrangement (solitary, generalized, annular, and linear), color (red, blue, brown, black, white, and yellow), border (well defined, poorly defined), shape(circular, strip, and irregular). Most of these styles can be found in Fig. 4. Other arrangement styles are also included in our dataset. In particular, the appearance diversity also exists in the same class, e.g. images in the first row of Fig. 4(a) contain skin disease images with different colors and shapes, coming from the same category named angioma.

(3) Attribute Diversity: Images in our dataset cover a lot of situations for patients such as age (child, adult, old), sex, disease site (hand, feet, head, nails), color of skin(white, yellow, brown, black) and different periods of lesions(early, middle, late). On the other hand, our dataset have also covered a lot of situations for environment, such as illumination, shooting distance, etc. All these diversities make our benchmark more comprehensive and challenging.

Challenge and Lack. Our dataset is a special images dataset different from object or scene datasets. The change of each condition, e.g. illumination, focal distance, and point of view, could increase a lot of difficulty for its classification. For example, the images with yellow boxes in the third row of Fig. 4(a) are from the same category named guttate psoriasis. The images from left to right are with different illumination and shooting distance leading to big differences among them. Furthermore, pathological changes in different periods and different

colors of skin of the patients all make a large intra-class variation. There are some diseases with low color contrast of foreground(lesion) and background(health skin), which are hard to recognize.

Of course, our dataset also has disadvantages. Details of dermatosis marks need stronger professional knowledge than other object and scene datasets. Considering the differences between clinical skin disease and fine-grained object images, e.g. birds and dogs, it's difficult for us to label part annotations in skin disease images. Besides, as Fig. 3 shows, our dataset shows imbalance among different categories. We try to collect the same number of samples, while some diseases rarely appear in real life.

4 Clinical Skin Disease Classification

In order to establish a baseline performance on our proposed dataset and evaluate the performance of different features, we design experiments for two aspects: (a) comparing the influence of different baseline features; (b) evaluating some existing methods whose aim is fine-grained classification. In all of the experiments, we randomly select half images from each class as the training set and the rest as testing set. We introduce our implementation details in the next paragraph. In addition, we present the color and texture features for classification and analyze their influences.

4.1 Hand-Crafted Features Based Classification

Implementation Details. We first investigate how conventional computer vision methods are used to recognize clinical skin disease images. We employ seven kinds of texture and color features and utilize LIBSVM, a popular library for support vector machine, to build some baseline algorithms. We use these algorithms to measure the classification accuracy on our dataset. Then, we evaluate our dataset using some existing work with their hand-tuned features and off-the-shelf frameworks. SIFT and Color Names features are extracted following the routine of [49]. HOG and LBP features are obtained by employing VLFeat [50].

Baseline Approaches. Two representative works [49,51] are included in this paper. Then, their performance on our proposed benchmark dataset are evaluated. While these methods are designed to classify fine-grained object or natural scene images, skin disease images are also sensitive to texture and color cues, which are employed in these tools.

In detail, Goering *et al.* [49] compute a global representation using the whole image. Feature types are the same as commonly used for fine-grained classification, i.e. bag-of-visual words with SIFT and Color Names, but with additional spatial pyramid pooling. Furthermore, they apply GrabCut segmentation to estimate the foreground. This algorithm performs iterative segmentation with a conditional Markov random field, where unary potentials are modeled with a Gaussian mixture model re-estimated in each iteration, and pairwise potentials

Table 3. Classification performance with different hand-engineered features on both of our datasets, i.e. SD-198 and SD-128. Each of the first seven methods is built with a popular off-the-shelf feature, using SVM as its classifier. On the other hand, the last two methods are designed for similar vision tasks, i.e. fine-grained object classification and natural scene classification, respectively.

Num	Features	Features dimension	Classifier	SD-198 %	SD-128 %
1	SIFT	21000	SVM	25.85	29.40
2	HOG	12400	SVM	12.78	14.17
3	LBP	23200	SVM	15.46	17.09
4	Color Histogram	768	SVM	4.19	5.59
5	Color Names(CN)	21000	SVM	20.20	20.32
6	Gist	512	SVM	16.49	17.52
7	Gabor	4000	SVM	10.14	11.37
Num	Methods	Features dimension	Methods or features	SD-198 %	SD-128 %
8	[49]	21000	SIFT+CN+SVM	52.19	53.29
9	[51]	4200	Spatial Pyramid	22.45	24.45

are added to favor strong image edges. Lazebnik *et al.* [51] have presented a holistic approach for image categorization based on a modification of pyramid match kernels. They repeatedly subdivide an image and compute histograms of image features over the resulting subregions, showing promising results on scene databases.

Results and Analysis. To establish a baseline performance on our dataset, we evaluate the features mentioned in Table 3. The experimental results show that texture features play a more important role than color features in this dataset. We find that the colors of foreground and background are extremely the same in some skin disease images. On the other hand, the lesions often present different textures and shapes, such as annular, linear, concave and convex shapes.

Furthermore, there are different skin disease categories sharing very similar shapes, and their color cures are slightly different, e.g. neurofibroma and apocrine hydrocystoma. Considering these cases, the off-the-shelf tool [49], performs best in this configuration, although it's designed for fine-grained object recognition. Note that, the influence of background clutter is significant in this method.

4.2 Deep Features Based Classification

Implementation Details. In our experiments, we extract deep convolutional features from a CNN model pre-trained on ImageNet. Due to the skin classes in our dataset, we change the original 1000-way fc8 classification layer to a new 198-way fc8 layer, whose weights are randomly initialized by a Gaussian function. We set fine-tuning learning rates as proposed by CaffeNet CNN, and initialize the global rate to a tenth of the initial ImageNet learning rate. In addition, during

Table 4. The average classification accuracy with different models of convolutional neural networks.ft indicates that the corresponding model is fine-tuned with our training samples

Method	SD-198[%]	SD-128[%]
CaffeNet	42.31	42.83
CaffeNe+ft	46.69	47.38
VGG	37.91	39.27
VGG+ft	50.27	52.15

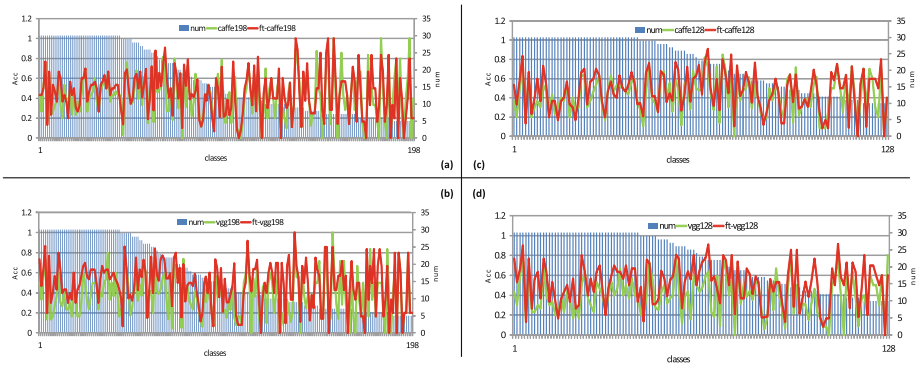

Fig. 5. Accuracy for each class with different models. (a) The performance of CaffeNet on SDC-198. (b) The performance of VGGNet on SDC-198. (c) The performance of CaffeNet on SDC-128. (d) The performance of VGGNet on SDC-128. For each figure, the secondary Y-axis(right) represents the number of testing image.

the training process, we drop the learning rate by a factor of 10. Furthermore, we independently fine-tune the ImageNet pre-trained CNN for classification on ground truth crops of each region warped to the 227×227 network input size. At test time, we extract features from the test images using the network fine-tuned on the training set of our skin disease images. Meanwhile, we also fine-tune a very deep CNN architecture, i.e. VGGNet [52] with 16 layers, to extract deep features.

Results and Analysis. We fine-tune the pre-trained CNN model, and compare it with the original CaffeNet by showing the results of using the SVM as a classifier. We extract deep features from the last layer of CaffeNet and obtain a 4096 dimensional feature representation. For both of our skin disease datasets, i.e. SD198 and SD-128, half images of each class are used for fine-tuning the model. From Table 4, we can draw a conclusion that the fine-tuned VGGNet gets significant promotion, which is mainly benefited from our lager-scale well-labeled dataset.

To further analyze their performance, we also calculate the accuracy for each class of the CNN models. Figure 5(a, b) show the classification results on SD-198, and Fig. 5(c, d) show the classification results on SD-128. It's shown that the accuracies have bigger fluctuation when the number of images of each class decreases. For these classes, the skin diseases have a relatively low morbidity in our daily life. Furthermore, we observe these classes, including stomatitis, histiocytosis-X, lymphangioma-circumscriptum, pomade-acne, etc., and find they share a common point that the corresponding images usual carry strong landmarks of lesions. For example, the region of skin disease is a saliency area. On the other hand, we find the classes has accuracy close to 0, which almost are hard to distinguish even for the professional doctor. For these classes, we may need more labeled data to provide in-sight to their characteristics.

5 Discussion

We have shown the performance of traditional features that have been commonly used in computer vision tasks. We also execute experiments with deep visual features on our skin disease benchmark dataset. The accuracy for all these features have been showed in Tables 3 and 4, respectively. In this section, we will compare the best performance of hand-crafted features with the deep visual features.

For SDC-198, the best classification result is 52.19 %, which is acquired by combining the SIFT and Color Names features. The accuracy using a pre-trained and fine-tuned VGGNet is 50.27 %. It is interesting to find that the performance

Fig. 6. Examples of classification results on our proposed benchmark dataset. (a) Images are correctly classified by [49] and wrongly classified by VGGNet. (b) Images are correctly classified by deep network and wrongly classified by [49] (Color figure online)

of hand-crafted features is better than deep visual features for the skin disease classification.

In order to investigate the reason, in Fig. 6(a), we show some representative images which have been correctly classified by [49] and wrongly classified by VGGNet. We also show the images in the opposite situation in Fig. 6(b). Useful observation can be draw from the presented images. First, images in (a) always have a cleaner background than the disease images in (b), and second, the appearance of lesions in (a) is much simpler than (b). Since [49] has applied a segmentation procedure with GrabCut to estimate the foreground, it's reasonable that this algorithm outperforms CNNs when both of them are applied to images in (a). For example, consider the images in the first row of both (a) and (b), these images are corresponding to skin diseases such as dermatofibroma, basal cell carcinoma, angioma, seborrheic keratosis and blue nevus etc. Compared to images in (a), the lesions in (b) are surrounded with more hair, which will weaken the segmentation employed in [49]. Moreover, CNNs have shown advantages in finding structure and semantic information. Images in (b) include more cues about the location of lesion, e.g. mouth, foot, eye, hand, etc., perform better with powerful VGGNet.

6 Conclusion

In this paper, we raise a challenging problem of automatic visual classification of clinical skin disease images. The absence of benchmark datasets is a barrier to a more dynamic development of this research area. We build a new and challenging clinical skin disease images dataset, including 6,584 real-world images from 198 categories. Each sample in our benchmark is well labeled. We intend to release the dataset to the community to promote the related research. Furthermore, we also evaluate the performance of different features to establish a baseline performance on our dataset.

Acknowledgements. This work was supported by the National Natural Science Foundation of China (No. 61301238, 61201424), China Scholarship Council (No. 2015 06205024) and the Natural Science Foundation of Tianjin, China (No.14ZCDZGX00831).

References

1. Hay, R.J., Johns, N.E., Williams, H.C., Bolliger, I.W., et al.: The global burden of skin disease in 2010: an analysis of the prevalence and impact of skin conditions. J. Invest. Dermatol. **134**(6), 1527–1534 (2014)
2. Lynch, P.J., Edwards, L.: Genital Dermatology. Churchill Livingstone, New York (1994)
3. Gonzalez-Castro, V., Debayle, J., Wazaefi, Y., Rahim, M., Gaudy-Marqueste, C., Grob, J.J., Fertil, B.: Automatic classification of skin lesions using color mathematical morphology-based texture descriptors. In: QCAV, p. 953409 (2015)

4. Badano, A., Revie, C., Casertano, A., Cheng, W.C., Green, P., Kimpe, T., Krupinski, E., Sisson, C., Skrøvseth, S., Treanor, D., et al.: Consistency and standardization of color in medical imaging: a consensus report. J. Digit. Imaging **28**(1), 41–52 (2015)
5. Glaister, J., Wong, A., Clausi, D., et al.: Segmentation of skin lesions from digital images using joint statistical texture distinctiveness. IEEE Trans. Biomed. Eng. **61**(4), 1220–1230 (2014)
6. Razeghi, O., Zhang, Q., Qiu, G.: Interactive skin condition recognition. In: ICME, pp. 1–6 (2013)
7. Yao, B., Khosla, A., Fei-Fei, L.: Combining randomization and discrimination for fine-grained image categorization. In: CVPR, pp. 1577–1584 (2011)
8. Wang, J., Yang, J., Yu, K., Lv, F., Huang, T., Gong, Y.: Locality-constrained linear coding for image classification. In: CVPR, pp. 3360–3367 (2010)
9. Branson, S., Wah, C., Schroff, F., Babenko, B., Welinder, P., Perona, P., Belongie, S.: Visual recognition with humans in the loop. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010. LNCS, vol. 6314, pp. 438–451. Springer, Heidelberg (2010). doi:10.1007/978-3-642-15561-1_32
10. Farrell, R., Oza, O., Zhang, N., Morariu, V., Darrell, T., Davis, L.S., et al.: Birdlets: Subordinate categorization using volumetric primitives and pose-normalized appearance. In: ICCV, pp. 161–168 (2011)
11. Chai, Y., Lempitsky, V., Zisserman, A.: Symbiotic segmentation and part localization for fine-grained categorization. In: ICCV, pp. 321–328 (2013)
12. Welinder, P., Branson, S., Mita, T., Wah, C., Schroff, F., Belongie, S., Perona, P.: Caltech-UCSD birds 200 (2010)
13. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The caltech-UCSD birds-200-2011 dataset (2011)
14. Nilsback, M.E., Zisserman, A.: Automated flower classification over a large number of classes. In: ICVGIP, pp. 722–729 (2008)
15. Angelova, A., Zhu, S., Lin, Y.: Image segmentation for large-scale subcategory flower recognition. In: WACV, pp. 39–45 (2013)
16. Kumar, N., Belhumeur, P.N., Biswas, A., Jacobs, D.W., Kress, W.J., Lopez, I.C., Soares, J.V.: Leafsnap: a computer vision system for automatic plant species identification. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part II. LNCS, vol. 7573, pp. 502–516. Springer, Heidelberg (2012)
17. Parkhi, O.M., Vedaldi, A., Zisserman, A., Jawahar, C.: Cats and dogs. In: CVPR, pp. 3498–3505 (2012)
18. Khosla, A., Jayadevaprakash, N., Yao, B., Li, F.F.: Novel dataset for fine-grained image categorization: Stanford dogs. In: CVPRW (2011)
19. Yang, L., Luo, P., Change Loy, C., Tang, X.: A large-scale car dataset for fine-grained categorization and verification. In: CVPR, June 2015
20. Duan, K., Parikh, D., Crandall, D., Grauman, K.: Discovering localized attributes for fine-grained recognition. In: CVPR, pp. 3474–3481 (2012)
21. Yao, B., Bradski, G., Fei-Fei, L.: A codebook-free and annotation-free approach for fine-grained image categorization. In: CVPR, pp. 3466–3473 (2012)
22. Yang, S., Bo, L., Wang, J., Shapiro, L.G.: Unsupervised template learning for fine-grained object recognition. In: NIPS, pp. 3122–3130 (2012)
23. Krause, J., Jin, H., Yang, J., Fei-Fei, L.: Fine-grained recognition without part annotations. In: CVPR (2015)
24. Wolff, K., Johnson, R.A., Suurmond, D.: Color Atlas and Synopsis of Clinical Dermatology. Mcgraw-hill Medical Pub. Division, New York (2005)

25. Maglogiannis, I., Delibasis, K.K.: Enhancing classification accuracy utilizing globules and dots features in digital dermoscopy. Comput. Methods Programs Biomed. **118**(2), 124–133 (2015)
26. Celebi, M.E., Iyatomi, H., Schaefer, G., Stoecker, W.V.: Lesion border detection in dermoscopy images. Comput. Med. Imaging Graph. **33**(2), 148–153 (2009)
27. Celebi, M.E., Kingravi, H.A., Uddin, B., Iyatomi, H., Aslandogan, Y.A., Stoecker, W.V., Moss, R.H.: A methodological approach to the classification of dermoscopy images. Comput. Med. Imaging Graph. **31**(6), 362–373 (2007)
28. Kasmi, R., Mokrani, K.: Classification of Malignant Melanoma and Benign Skin Lesions: implementation of automatic ABCD rule. IET Image Process. **10**(6), 448–455 (2016)
29. Argenziano, G., Soyer, H.P., De Giorgi, V., Piccolo, D., Carli, P., Delfino, M., et al.: Interactive Atlas of Dermoscopy (Book and CD-ROM), pp. 1–10. EDRA Medical Publishing & New Media (2000)
30. Abbas, Q., Fondón, I., Rashid, M.: Unsupervised skin lesions border detection via two-dimensional image analysis. Comput. Methods Programs Biomed. **104**(3), 1–15 (2011)
31. Wazaefi, Y., Paris, S., Fertil, B.: Contribution of a classifier of skin lesions to the dermatologist's decision. In: IPTA, pp. 207–211 (2012)
32. Sadeghi, M., Lee, T.K., McLean, D., Lui, H., Atkins, M.S.: Detection and analysis of irregular streaks in dermoscopic images of skin lesions. IEEE Trans. Med. Imaging **32**(5), 849–861 (2013)
33. Barata, C., Emre Celebi, M., Marques, J.S.: Melanoma detection algorithm based on feature fusion. In: EMBC, pp. 2653–2656 (2015)
34. Mendonça, T., Ferreira, P.M., Marques, J.S., Marcal, A.R., Rozeira, J.: PH 2-A dermoscopic image database for research and benchmarking. In: EMBC, pp. 5437–5440 (2013)
35. Glaister, J.L.: Automatic segmentation of skin lesions from dermatological photographs. MSc Dissertation, Dept. Systems Design Eng., University of Waterloo, Ontario, Canada (2013)
36. Cho, D.S., Haider, S., Amelard, R., Wong, A., Clausi, D.A.: Quantitative features for computer-aided melanoma classification using spatial heterogeneity of eumelanin and pheomelanin concentrations. In: ISBI, pp. 59–62 (2015)
37. Razeghi, O., Qiu, G., Williams, H., Thomas, K.: Skin lesion image recognition with computer vision and human in the loop. In: Medical Image Understanding and Analysis (MIUA), Swansea, UK, pp. 167–172 (2012)
38. Alcón, J.F., Ciuhu, C., Ten Kate, W., Heinrich, A., Uzunbajakava, N., Krekels, G., Siem, D., De Haan, G.: Automatic imaging system with decision support for inspection of pigmented skin lesions and melanoma diagnosis. IEEE J. Sel. Top. Sign. Proces. **3**(1), 14–25 (2009)
39. Razeghi, O., Fu, H., Qiu, G.: Building skin condition recogniser using crowdsourced high level knowledge. In: Medical Image Understanding and Analysis (MIUA), Birmingham, UK, pp. 225–230 (2013)
40. Razeghi, O., Qiu, G.: 2309 skin conditions and crowd-sourced high-level knowledge dataset for building a computer aided diagnosis system. In: ISBI, pp. 61–64 (2014)
41. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: NIPS, pp. 1097–1105 (2012)
42. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: CVPR, pp. 580–587 (2014)
43. Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., Oliva, A.: Learning deep features for scene recognition using places database. In: NIPS, pp. 487–495 (2014)

44. Xiao, T., Xu, Y., Yang, K., Zhang, J., Peng, Y., Zhang, Z.: The application of two-level attention models in deep convolutional neural network for fine-grained image classification. In: CVPR, pp. 842–850 (2014)
45. Zhang, N., Donahue, J., Girshick, R., Darrell, T.: Part-based R-CNNs for fine-grained category detection. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8689, pp. 834–849. Springer, Heidelberg (2014). doi:10.1007/978-3-319-10590-1_54
46. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: ICLR (2015)
47. Marks, J.G., Miller, J.J.: Lookingbill and Marks' Principles of Dermatology, pp. 7–10. Elsevier Health Sciences (2013)
48. Friedman, R.J., Rigel, D.S., Kopf, A.W.: Early detection of malignant melanoma: the role of physician examination and self-examination of the skin. CA Cancer J. Clin. **35**(3), 130–151 (1985)
49. Goering, C., Rodner, E., Freytag, A., Denzler, J.: Nonparametric part transfer for fine-grained recognition. In: CVPR, pp. 2489–2496 (2014)
50. Vedaldi, A., Fulkerson, B.: VLFeat: an open and portable library of computer vision algorithms (2008). http://www.vlfeat.org/
51. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In: CVPR, vol. 2, pp. 2169–2178 (2006)
52. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. CoRR abs/1409.1556 (2014)