HYBRID DERMOSCOPY IMAGE CLASSIFICATION FRAMEWORK BASED ON DEEP CONVOLUTIONAL NEURAL NETWORK AND FISHER VECTOR

Zhen Yu[1], *Dong Ni*[1], *Siping Chen*[1], *Jin Qin*[2], *Shengli Li*[3], *Tianfu Wang*[1*],*and Baiying Lei*[1*]
[1]School of Biomedical Engineering, Shenzhen University,
National-Regional Key Technology Engineering Laboratory for Medical Ultrasound,
Guangdong Key Laboratory for Biomedical Measurements and Ultrasound Imaging, Shenzhen, China.
[2]Centre for Smart Health, School of Nursing, The Hong Kong Polytechnic University, Hong Kong
[3]Department of Ultrasound, Affiliated Shenzhen Maternal and Child Healthcare
Hospital of Nanfang Medical University, Shenzhen, P.R.China.
(*Email:{ tfwang,leiby}@szu.edu.cn)

ABSTRACT

Dermoscopy image is usually used in early diagnosis of malignant melanoma. The diagnosis accuracy by visual inspection is highly relied on the dermatologist's clinical experience. Due to the inaccuracy, subjectivity, and poor reproducibility of human judgement, an automatic recognition algorithm of dermoscopy image is highly desired. In this work, we present a hybrid classification framework for dermoscopy image assessment by combining deep convolutional neural network (CNN), Fisher vector (FV) and support vector machine (SVM). Specifically, the deep representations of subimages at various locations of a rescaled dermoscopy image are first extracted via a natural image dataset pre-trained on CNN. Then we adopt an orderless visual statistics based FV encoding methods to aggregate these features to build more invariant representations. Finally, the FV encoded representations are classified for diagnosis using a linear SVM. Compared with traditional low-level visual features based recognition approaches, our scheme is simpler and requires no complex preprocessing. Furthermore, the orderless representations are less sensitive to geometric deformation. We evaluate our proposed method on the ISBI 2016 Skin lesion challenge dataset and promising results are obtained. Also, we achieve consistent improvement in accuracy even without fine-tuning the CNN.

Index Terms— Dermoscopy image, Fisher vector, Deep convolutional neural network, Classification

1. INTRODUCTION

Melanoma skin cancers is one of the most rapidly increasing and deadliest cancers in the world, which accounting for 79% of skin cancer deaths [1-3]. Early diagnosis is of great importance for treating this disease as it can be cured more easily at early stages [1-4]. To improve the diagnosis of this disease, dermoscopy has been introduced to assist dermatologists in clinical examination since it is a non-invasive skin imaging technique that provides clinicians high quality visual perception of skin lesion. Compared with the conventional macroscopic (clinical) images, fewer surface reflection, sufficient deeper layers' details, and lower screening errors play an important role in visibility and recognition accuracy of dermoscopy image [2, 5]. Since melanoma is more deadly than non-melanoma skin cancer (examples shown in Fig. 1), discrimination between cancer and non-cancerous melanoma dermoscopy images has attracted considerable interest [1, 3, 4]. Clinically, several heuristic approaches, such as "ABCD" rule [6], Menzies method [7] and CASH [8], have been developed to enhance clinicians' ability to distinguish melanomas from benign nevi. However, the correct diagnosis of a skin lesion is not trivial even for health care professionals. Furthermore, dermoscopic diagnosis made by human visual inspection is often subjective. Hence, unsatisfactory accuracy and poor reproducibility are still an issue for diagnosing this disease.

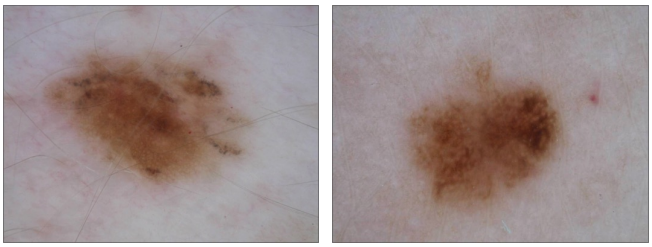

Fig. 1. Example dermoscopy image of skin lesions.

To tackle these issues, numerous automatic algorithms were proposed for dermoscopic image analysis [1]. Interested readers can refer to [2, 9] for a comprehensive summary of related work over the past decades. In general, existing approaches vary from medical image, medicine to computer vision domain. Within the domain of computer vision, existing methods mainly focus on combining low-level hand-crafted features (e.g., RGB statistical descriptors [10], lesion's area & moment of inertia [11]) with machine learning models (e.g., support vector machine (SVM), and artificial neural network [3]). Also, there are some detection and segmentation work [3]. However, due to the high intra-

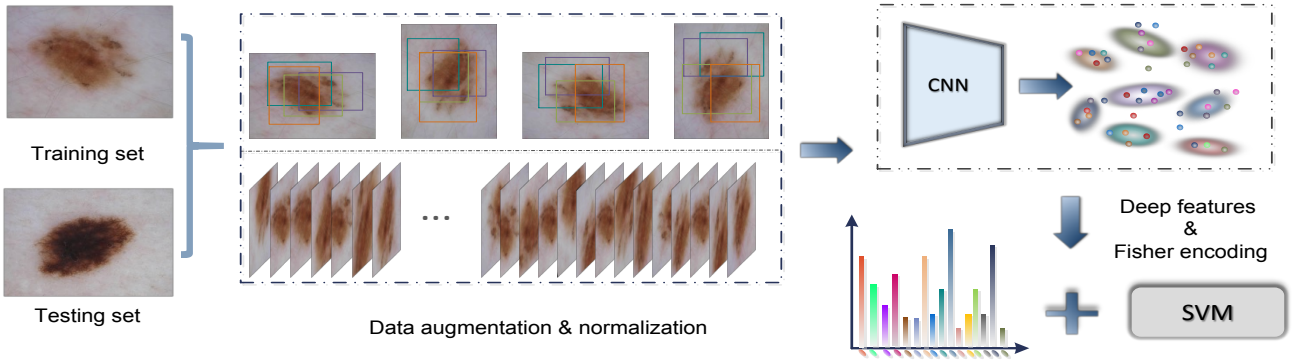

Fig. 2. Flowchart of proposed method.

class and low inter-class variations of melanoma, low-level visual features based diagnostic performance is still unsatisfactory.

Differing from the low-level feature representation, deep learning methods such as deep convolutional neural network (CNN) have witnessed a great success in image recognition tasks in the past few years. The main reason is that CNN is endowed with an impressive representation capability for the recognition or detection task depending on the given training dataset. Also, CNN model contains multiple processing layers to learn different level features. Hence, combining these hierarchical features preserves extremely discriminative and effective deep representations. State-of-the-art performance was achieved in numerous applications [12, 13]. Although CNN produces attractive results when trained for specific task, study in [14] demonstrated that CNN architectures pre-trained on large ImageNet dataset [15] also delivered promising results for other image tasks without retraining. For this reason, CNN has been widely applied in dermoscopy image classification. For example, Codella *et al*. [4] adopted ImageNet pre-trained CNN to extract high level feature representations to differentiate melanoma and non-melanoma images. Kawahara *et al*. [16] used pre-trained CNN as a feature extractor and combined subimages features pooling for 10-classes lesion classification. Demyanov *et al*. [17] developed a five layer CNN architecture to differentiate two types of skin lesion data. Although high dimensional CNN based features have good generalization of image representation, the deep descriptors are sensitive and vulnerable to geometric variations [18]. For images with various scales, it poses a great challenge to the classification directly using CNN features. To solve it, rescaling and data augmentation (crop, flip or rotate) strategies are commonly used solutions [14]. However, the improvements in geometric invariance is highly limited. To overcome the limitations, we integrate the Fisher vector (FV) encoding strategy with deep features in our task since FV encoding of low level descriptor has obtained impressive performance in image recognition [19-21].

In this paper, we present a hybrid framework that combines CNN features extracted from multiple subimages with robust and discriminative feature representation encoded by FV for dermoscopy images classification. Instead of training the entire CNN and adopting preprocessing such as lesion detection and segmentation, our scheme combines the advantages of both deep features and local descriptors encoding methods efficiently. We evaluate our approach using the public ISBI 2016 skin lesion classification data [5]. Our experimental results indicate that the proposed method achieves quite promising accuracy and outperforms the state-of-art methods as well.

2. METHODOLOGY

Generally, the pipeline of existing skin lesion image classification works can be summarized as follows: 1) preprocess the images, such as hair removal [22], image enhancement; 2) border detection or segmentation; 3) feature extraction; 4) classification. Fig.2 illustrates the flowchart of our framework. In fact, the main idea of our proposed method is to sample a set of subimages from each rescaled image, and apply the pre-trained CNN over orderless patches to generate local deep representations. In addition, Fisher encoding is adopted to aggregate these local descriptors in the final image representation.

2.1. Image preprocessing and augmentation

There is a huge variation in image sizes of the Skin lesion dermoscopy images dataset, ranging from the largest scale (4288×2848) to the smallest scale (722×542). For this reason, we re-scale these images along the shortest side to a uniform size while maintaining the aspect ratio. For each scaled image, we rotate them in four degrees (0, 90, 180, and 270). Also, sampling windows with various sizes are generated and each window captures the local areas to construct a subimage. These subimages (augmented images) are then wrapped into a fixed 227×227 size. Before feeding into CNN, images are normalized by subtracting the mean pixel value, calculated over the entire training dataset so that the RGB values are centered at zero (denoted as all-img-mean). As indicated in [16], normalizing the skin lesion images with mean pixel over individual image (denoted as per-img-mean) provides additional improvements in

accuracy. In this work, we investigate the influence of two different normalization approaches.

2.2. Deep feature representation

To obtain image representations from CNN, responses of the intermediate layers (i.e. fully-connected layers) are usually extracted and treated as general image features by feeding this image into the network. In our scheme, we represent a skin lesion image as multiple subimages, and extract deep feature for these subimages. Specifically, we adopt an eight layers CNN (AlexNet) [13] pre-trained on the large ImageNet. The network takes images of 227×227 size as input [13]. For a skin lesion image, as described in Section 2.1, we randomly extract hundreds of subimages. Before passing these subimages through the network, these local patches with various scales are wrapped into the required size without considering their aspect ratio. The output of the first fully connected layer (FC6) is obtained as deep representation of current input patch. Thus, we obtain a set of 4096-dimensional feature vectors for each dermoscopy image. We also investigate the performance when higher layer (FC7) output is adopted. It is noteworthy that we focus our study on the combination of deep features and Fisher encoding for image representation and classification. Hence, we do not consider the classification performance variance among different architectures.

2.3. Fisher encoding

In our study, we consider deep features of augmented images of a skin lesion image as generic local descriptors, and then perform FV encoding over these deep visual descriptors. Specifically, given an *i*-th skin image χ^i, after the preprocessing and augmentation, we have a set of augmented images ,$\{x_1^i, x_2^i \dots x_N^i\}$. We normalize the augmented images and input them into a pre-trained CNN. Hence, a total of 4096-dimensional *N* feature vectors is then obtained $\mathcal{F}^i = \{f_1^i, f_2^i \dots f_N^i\}$. Note that *N* is fixed to 256 in our experiment. Each feature vector is processed by l_2 normalization. For more efficient computation, the feature vectors are further reduced by principle component analysis (PCA) to dimensionality of 128. Similar to [21], Gaussian mixture model (GMM) with *K* components is first learned by sampled images from training set. FV representation for the current image (the *i*-th image) is calculated as follow:

$$u_k = \frac{1}{N\sqrt{\pi_k}} \sum_{j=1}^{N} q_{kj} \left(\frac{f_j^i - \mu_k}{\sigma_k} \right), \quad (1)$$

$$v_k = \frac{1}{N\sqrt{2\pi_k}} \sum_{j=1}^{N} q_{kj} \left[\left(\frac{f_j^i - \mu_k}{\sigma_k} \right)^2 - 1 \right], k=1, 2 \cdots K \quad (2)$$

$$\Phi_i = [\, u_1, v_1, \dots u_K, v_K], \quad (3)$$

where μ_k, σ_k are the mean and covariance of the GMM model for each GMM cluster, respectively, π_k is the parameters of GMM model, q_{kj} is the soft-assignment of each feature [23]. By concatenating u_k and v_k for all *K* components, we obtain the final FV representation Φ_i, which is used for the classifier learning.

3. EXPERIMENTS AND RESULTS

3.1. Experiment setting

In this study, we validate our proposed method using the ISBI 2016 skin lesion classification data [5]. This dataset contains 1279 skin images with the corresponding class labels, which are pre-partitioned into a training set of 900 images and a testing set of 379 images. There are two lesion categories in the dataset: Melanoma and Benign (Non-Melanoma). Apart from our proposed method, we also compare several existing approaches. For example, extracting CNN features per image or multiple augmented images and performing classification using SVM, we denote them as CNN-SVM and CNNaug-SVM, respectively. AVG-Pooling represents the method of average pooling the deep features of multiple augmented images. For performance metrics, we adopt the mean average precision (mAP), accuracy (Acc) and area under receive operation curve (AUC) to evaluate the performance of each approach.

3.2. Results

We start by investigating the influence of normalized images before feeding into CNN. We carry out the experiment with different rescaled images and perform two normalizations described in Section 2 (per-img-mean and all-img-mean). The experimental results are shown in Table 1. We can see that adopting per-img-mean normalization has slight performance improvements in mAP. Hence, in the following experiments, all images are normalized by per-img-mean. For different layers of CNN features, outputs of FC6 and FC7 are extracted and encoded by FV, then linear SVM are trained for classification. The results demonstrate that FC6 features perform better than FC7 features, and there is around 4% performance improvement in terms of mAP. In order to further demonstrate the geometric invariance of FV representations, we conduct experiments without FV encoding (CNNaug-SVM) in different scales. The comparison results are illustrated in Fig. 3.

Finally, we summarize the algorithm comparison results of our proposed method and other CNN based methods in Table 2. In CNNaug-SVM case, we extract the same number of subimages as our method (CNN-FV), and then average the scores. For the fusion method, we concatenate the FV representations of an image in four different scales, and the resulting representation is further compressed to 32768 dimensionality for more efficiency. We observe the combination of multi-scale images and FV representations leads to improvements of ~3%, ~1%, and ~3% margins in mAP, Acc and AUC, respectively, which demonstrates the effectiveness of our proposed method. Overall, better results is achieved by integrating state-of-art deep CNN architecture with discriminative FV encoding method.

Table 1. Classification results of proposed approach (We keep number of Gaussian as 128 in all experiments. For each image, FV representation dimension is 32768).

Model	Normalization	Scale (short side)	mAP (%)	Acc (%)	AUC (%)
CNN-FC6 + FV	all-img-mean	256	47.00	80.21	72.84
	all-img-mean	355	48.27	81.53	73.15
CNN-FC6 + FV	per-img-mean	256	**50.23**	**82.32**	**76.89**
	per-img-mean	355	50.07	81.79	78.39
	per-img-mean	512	46.58	81.00	76.40
CNN-FC7 + FV	per-img-mean	256	44.17	80.47	70.89
	per-img-mean	355	49.56	82.85	77.24
	per-img-mean	512	39.03	79.68	73.16

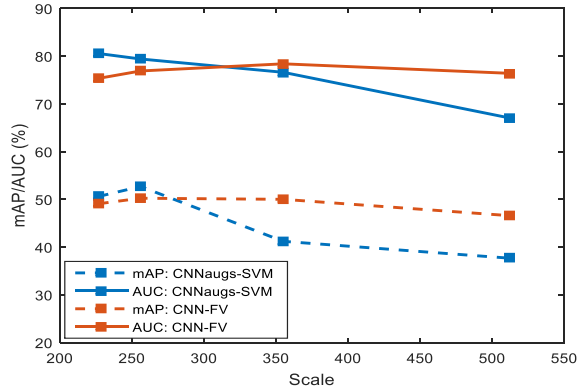

Fig. 3. Comparison of CNNaug-SVM and CNN-FV for different scales.

Table 2. Classification results compared with existing CNN-based methods.

Model	Feature	Dim	mAP (%)	Acc (%)	AUC (%)
CNN-SVM	AlexNet-FC6	4096	43.12	72.30	74.45
CNNaug-SVM [14]	AlexNet-FC6	4096	52.47	79.40	76.78
AVG-Pooling	AlexNet-FC6	4096	39.12	75.21	68.65
VLAD-SVM [18]	AlexNet-FC6	51200	49.09	81.53	77.82
Full-trianed CNN [13]	---	---	48.41	82.59	75.32
Proposed	AlexNet-FC6	32768	50.23	82.32	76.89
Proposed (fusion)	AlexNet-FC6	32768	**53.50**	**83.09**	**79.57**

4. CONCLUSION

In this paper, we investigated a hybrid framework for dermoscopy images classification. Our experiments showed that deep features extracted by pre-trained CNN can be efficiently aggregated by FV. The encoded deep representations possess better scale invariance than general CNN features. Hence, they are more suitable for holistic images representation. Furthermore, our proposed method achieves promising classification result by combining multi-scale image information. Our future work will focus on studying deep features of different CNN architectures (e.g. ResNet) and multi-level features within a network.

5. REFERENCES

[1] R. Kasmi and K. Mokrani, "Classification of malignant melanoma and benign skin lesions: implementation of automatic ABCD rule," *IET Image Processing*, vol. 10, no. 6, pp. 448-455, 2016.

[2] A. R. A. Ali and T. M. Deserno, "A systematic review of automated melanoma detection in dermatoscopic images and its ground truth data," *Proc. SPIE Med. Imag.*, vol. 8318, no. 1, pp. 83181I-1-83181I-11, 2012.

[3] K. H. M. Celebi, B. Uddin, H. Iyatomi, Y. Aslandogan, W. Stoecker, R. Moss, "A methodological approach to the classification of dermoscopy images," *Comput. Med. Imag. Grap.*, vol. 31, no. 6, pp. 362-373, 2007.

[4] N. Codella, J. Cai, M. Abedini, R. Garnavi, A. Halpern, and J. R. Smith, "Deep Learning, Sparse Coding, and SVM for Melanoma Recognition in Dermoscopy Images," in *MICCAI*, 2015, pp. 118-126.

[5] D. Gutman, N. C. Codella, E. Celebi, B. Helba, M. Marchetti, N. Mishra, and A. Halpern, "Skin Lesion Analysis toward Melanoma Detection: A Challenge at the International Symposium on Biomedical Imaging (ISBI) 2016, hosted by the International Skin Imaging Collaboration (ISIC)," *arXiv: 1605.01397*, 2016.

[6] W. Stolz, A. Riemann, A. Cognetta, L. Pillet, W. Abmayr, D. Holzel, P. Bilek, F. Nachbar, and M. Landthaler, "Abcd rule of dermatoscopy-a new practical method for early recognition of malignant-melanoma," *Eur. J. Dermatol.*, vol. 4, no. 7, pp. 521-527, 1994.

[7] S. W. Menzies, C. Ingvar, K. A. Crotty, and W. H. McCarthy, "Frequency and morphologic characteristics of invasive melanomas lacking specific surface microscopic features," *Arch. Dermatol.*, vol. 132, no. 10, pp. 1178-1182, 1996.

[8] J. S. Henning, S. W. Dusza, S. Q. Wang, A. A. Marghoob, H. S. Rabinovitz, D. Polsky, and A. W. Kopf, "The CASH (color, architecture, symmetry, and homogeneity) algorithm for dermoscopy," *J. Amer. Acad. Dermatol.*, vol. 56, no. 1, pp. 45-52, 2007.

[9] K. Korotkov and R. Garcia, "Computerized analysis of pigmented skin lesions: A review," *Artif. Intell. Med.*, vol. 56, no. 2, pp. 69-90, 2012.

[10] H. Iyatomi, K. A. Norton, M. E. Celebi, G. Schaefer, M. Tanaka, and K. Ogawa, "Classification of melanocytic skin lesions from non-melanocytic lesions," in *EMBC*, 2010, pp. 5407-5410.

[11] Z. She, Y. Liu, and A. Damatoa, "Combination of features from skin pattern and ABCD analysis for lesion classification," *Skin Res. Technol.*, vol. 13, no. 1, pp. 25-33, 2007.

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *arXiv: 1512.03385*, 2015.

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inform. Process. Syst.*, 2012, pp. 1097-1105.

[14] A. Sharif Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN features off-the-shelf: an astounding baseline for recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. DeepVision Workshop*, 2014, pp. 806-813.

[15] J. Deng, W. Dong, R. Socher, L. J. Li, L. Kai, and F.-F. Li, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248-255.

[16] J. Kawahara, A. BenTaieb, and G. Hamarneh, "Deep features to classify skin lesions," in *ISBI*, 2016, pp. 1397-1400.

[17] S. Demyanov, R. Chakravorty, M. Abedini, A. Halpern, and R. Garnavi, "Classification of dermoscopy patterns using deep convolutional neural networks," in *ISBI*, 2016, pp. 364-368.

[18] Y. Gong, L. Wang, R. Guo, and S. Lazebnik, "Multi-scale Orderless Pooling of Deep Convolutional Activation Features," in *ECCV*, 2014, pp. 392-407.

[19] B. Lei, W. Li, Y. Yao, X. Jiang, E. Tan, J. Qin, S. Chen, D. Ni, and T. Wang, "Multi-modal and multi-layout discriminative learning for placental maturity staging," *Pattern Recognit.* DOI: http://dx.doi.org/10.1016/j.patcog.2016.09.037, 2016.

[20] B. Lei, Y. Yao, S. Chen, S. Li, W. Li, D. Ni, and T. Wang, "Discriminative learning for automatic staging of placental maturity via multi-layer fisher vector," *Sci. Rep.*, vol. 5, no. p. 12818, 2015.

[21] J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek, "Image Classification with the Fisher Vector: Theory and Practice," *Int. J. Comput. Vision*, vol. 105, no. 3, pp. 222-245, 2013.

[22] T. Lee, V. Ng, R. Gallagher, A. Coldman, and D. McLean, "Dullrazor®: A software approach to hair removal from images," *Comput. Biol. Med.*, vol. 27, no. 6, pp. 533-543, 1997.

[23] D. Reynolds, "Gaussian mixture models," 2009, pp. 659-663.