Extension of automated melanoma screening for non-melanocytic skin lesions

Kouhei Shimizu and Hitoshi Iyatomi
Applied Informatics
Hosei University
Tokyo, Japan

Kerri-Ann Norton
Biomedical Engineering
Johns Hopkins University
Baltimore, MD, USA

M.Emre Celebi
Computer Science
Louisiana State University in Shreveport
Shreveport, LA, USA

***Abstract*—In this paper, we present an automated melanoma screening system that supports not only melanocytic skin lesions (MSLs) but also non-melanocytic skin lesions (NoMSLs). Melanoma is known as the most fatal skin cancer. Therefore, early detection is highly desired. However, melanoma diagnosis is not easy even for expert dermatologists. In such a background, several researchers have developed automated methods for melanoma detection but they mostly focused only on MSLs while NoMSLs have been almost neglected. To expand the scope to NoMSLs, we developed two melanoma classification models, namely the single-shot and the double-shot. The single-shot model differentiates melanomas from all the other skin lesions including NoMSLs. The double-shot model divides the task into two subtasks. Firstly, it differentiates MSLs from NoMSLs and then differentiates melanomas from the other MSLs. The single-shot achieved a sensitivity (SE) of 92.9% and a specificity (SP) of 83.9%, while the double-shot achieved an SE of 97.6% and an SP of 92.2% when 10 image features were used. The double-shot showed superior detection performance to the single-shot except when their constituent image features were limited.**

***Index Terms*—dermoscopy, melanoma, computer-aided diagnosis**

I. Introduction

Skin cancers have varying malignancy and symptoms. The three most common skin cancers are melanoma, basal cell carcinoma (BCC), and squamous cell carcinomas (SCC). Melanoma originates from melanocytes. It is the most life-threatening skin cancer because it grows fast and metastasizes rapidly. BCC originates from basal cells is the least harmful but has the highest incidence among all the skin cancers [1]. SCC originates from squamous cells grows slower than melanomas but develops metastasis at later stages.

Skin lesions that originate from melanocytes such as melanomas, Clark nevi, Spits nevi are called melanocytic skin lesions (MSLs). Other skin lesions originate not from melanocytes such as BCCs, SCCs, hematomas, hemangiomas, Seborrheic Keratoses are called as non-melanocytic skin lesions (NoMSLs). BCC and SCC account for the majority of skin cancers: BCC 80%, SCC 16%, and melanoma 4% in the United States [2]. Despite the low incidence of melanomas, 75% of death of overall skin cancers is caused by them [3].

Melanoma is especially difficult to cure after the metastasis stage. The five year survival rate is only 9-15% at the stage IV, while on the other hand, it goes up to 85-99% if detected at the stage Ⅱ [4]. Therefore, early detection is highly needed to reduce the mortality of melanoma [5]. Dermoscopy, a dermatological device to inspect skin lesions, is used for accurate melanoma diagnosis. With the aid of dermoscopy, accuracy in melanoma diagnosis has increased by 10-27% [6] but it was still remained around 75-88% [7]. To tackle the issues, several computer-aided melanoma detection systems have been developed [8]-[12]. The computer-aided techniques are expected to be a supplementary help for both inexperienced and experienced clinicians. The basic procedures for the automatic melanoma detection consists of the 3 steps: (1) tumor extraction, (2) feature extraction, and (3) classifier development.

As for some recent achievements, Celebi et al. [10] achieved a sensitivity (SE) of 93.3% and a specificity (SP) of 92.3% based on 564 dermoscopy images. Even though there are several limitations, this is one of promising results. We have been developing an Internet-based melanoma screening system [11] and keep investigating on improvement (current URL is http://dermoscopy.k.hosei.ac.jp). Our latest system was built based on a total of 1,455 dermoscopy images with the confirmed diagnosis and supports not only usual pigmented skin lesions, but also acral volar lesions [12]. The classification accuracy of the system is around 86% in SE and SP for the former and 93% in SE and 91% in SP for the latter. With the advantage of the Internet-connection, everyone who has a dermoscopy can use our system.

These conventional researches aimed at detecting melanomas from only MSLs while NoMSLs were almost neglected [13]. Therefore, if an NoMSL is fed into these systems, they may yield undesired results. One reason for NoMSLs being neglected is because discriminating melanomas from NoMSLs is comparatively easier than those from the other MSLs for dermatological experts. However, this differentiation is not always easy for physicians with different expertise and inexperienced dermatologists. When we consider to widen the target user of our Internet-based system also for those potential users, supporting NoMSLs is necessary.

In our previous studies, we have developed sophisticated tumor extraction algorithm for both MSLs and NoMSLs [13] and built a classifier for discriminating them [14]. In these studies, we confirmed that only two image features (“the skewness of bright region in the tumor along its major axis”

and “the difference between the average intensity in the peripheral part of the tumor and that in the normal skin area”) differentiated MSLs and NoMSLs quite well (98.0% in SE, 86.6% in SP).

In this paper, we describe the development of melanoma classification method that supports both MSLs and NoMSLs. The motivation behind the study is the strong demand of melanoma detection in early stages, and the desire to make the current system more practical also for physicians who have different expertise.

II. Material

We collected dermoscopy images with established diagnosis from Keio University Hospital in Japan. The following 3 sets of images were used in this study.

Dataset-A: 42 malignant melanomas (MSL-m): The primary concern of our study is to detect melanomas from all other skin lesions.

Dataset-B: 506 melanocytic nevi (MSL-n): This type of benign skin lesions has been the main target of automated melanoma diagnosis because of the difficulty in distinguishing them from melanomas.

Dataset-C: 110 non-melanocytic skin lesions (NoMSLs): 34 BCC, 59 Seborrheic Keratoses, 3 hemangioma, and 11 hematoma.

The datasets were 24-bit JPEG images with a typical resolution of 1136x852 pixels. The diagnosis was made by histopathological examination or clinical agreement by several expert dermatologists.

III. Method

A. Tumor Extraction

Analysis of border regions of the tumor is important for accurate diagnosis of skin lesions. Therefore, we need to segment out appropriate tumor area from dermoscopy images first. Conventional tumor extraction methods mostly focused on only MSLs whereas NoMSLs were neglected. This is because classification tasks for NoMSLs have got less attention, as mentioned before, and NoMSLs have a wide variety in appearance compared to MSLs and often have ambiguous borders. These properties make it difficult to segment out appropriate tumor area.

Recently, we developed a general tumor extraction algorithm that supports both MSLs and NoMSLs [13]. The method uses some color thresholding and morphology operations and outperforms all the other state-of-the-art methods for NoMSLs and equivalent or better for MSLs. We used this algorithm to segment out the tumor area for all the images in the datasets. Please refer to our original literature [13] for more detail.

B. Feature Extraction

After extracting the tumor areas, we calculated 428 image features [11] for each image with reference to the ABCD rule [15], commonly used clinical findings. The image features were designed to be scale-invariant and rotation-invariant. The breakdown of the features are: 80 asymmetry, 32 border, 140 color, and 176 texture.

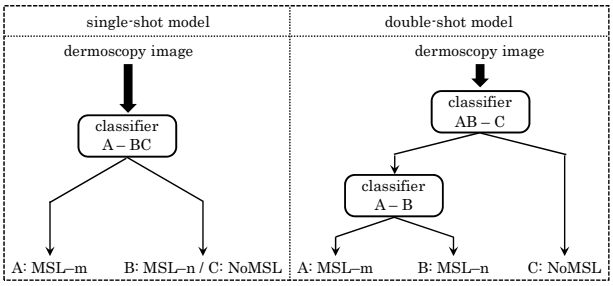

Fig. 1. Classification models

As for asymmetry features, circularity and gravity center of each brightness level etc. were calculated. For border features, color gradient between internal and peripheral of the tumor were calculated. For color features, basic color characteristics (e.g. minimum, average etc.) of the tumor (e.g. internal, peripheral, outside of the tumor) in several color channels (e.g. RGB channel, HSV channel) were calculated. In addition, the pleochroic features were also calculated. For texture features (represent “Dermoscopy structures” in the ABCD rule), correlation, energy, entropy, and moment features with different angle and unit resolution were calculated. Please refer to our previous literature [11] for detail.

C. Classification

We introduced two classification models for detecting melanomas: namely, the single-shot model and the double-shot model as illustrated in Fig. 1.

The single-shot determines whether the input is melanoma (A: MSL-m) or not (B: MSL-n, C: NoMSLs) by a single linear classifier “A-BC”. The double-shot detects melanoma with two steps. Firstly, it determines whether the input is MSLs (A, B) or NoMSLs (C) by the “AB-C”. If the input was classified as an MSL, the second classifier “A-B” determines whether the input is MSL-m (A) or MSL-n (B). The idea is to decompose the whole task into the two subtasks: (1) discard NoMSLs and keep MSLs, (2) discard MSL-n and detect MSL-m.

We used linear classifiers in both models. One of the most important steps for developing a classifier is to select appropriate features. We used an incremental stepwise method with a hypothesis test of Wilks’ lambda [16] in this study.

As for supervised output, we assigned +1 to MSL-m in the A-BC, MSLs in the AB-C, and MSL-m in the A-B classifiers. On the other hand, -1 was assigned to the other cases.

After the linear model is developed, the threshold for binary classification is adjusted to optimize the evaluation criteria such as sensitivity (SE) and specificity (SP). In this study, SE denotes the ratio of melanomas (A: MSL-m) successfully classified as melanomas. SP denotes the ratio of non-melanomas (B: MSL-n and C: NoMSLs) classified as either MSL-n or NoMSLs. Some erroneous cases such as MSL-n (B) is misclassified as NoMSLs (C) and vice-versa are also counted as successful classification because in melanoma screening, differentiation between MSL-n and NoMSLs is not

TABLE I
SELECTED FEATURES FOR EACH LINEAR CLASSIFIER

classifier	single-shot model		double-shot model			
	A-BC		AB-C		A-B	
	category	detail	category	detail	category	detail
Top 10 features	color	std-R (P) †	asymmetry	skew-x (180)	color	#HSV16 (T)
	color	ave-R (P-N) †	color	ave-B (P-T)	color	max-R (T-N)
	border	grad-V (1/20) ◇	color	ave-G (P-N)	asymmetry	std-y (55)
	color	std-H (P)	asymmetry	skew-y (155)	color	ave-S (N)
	asymmetry	std-x (255)	border	grad-B (1/5)	color	ave-R (P-N)
	color	#RGB16 (T) ⋆	asymmetry	skew-y (130)	color	skew-R (T-N)
	asymmetry	skew-x (205) ∗	asymmetry	std-x (180)	color	min-R (T-N)
	texture	correlation-90° (1/4) ‡	asymmetry	std-x (30)	color	ave-B (P-T)
	texture	correlation-135° (1/5.7)	color	min-B (P)	texture	correlation-90° (1/5.7)
	texture	correlation-45° (1/11.3)	asymmetry	circularity (30)	texture	correlation-135° (1/5.7)
#features	17		33		34	

The following abbreviations are used: target area: T(tumor), N(normal skin), P(peripheral); color channel: R/G/B and H/S/V; image axis: x(major axis of the tumor), y(minor axis of the tumor).
†: std-R (P): Standard deviation(std) of red(R) in peripheral(P) area; ave-R (P-N): Difference in average of red(R) between peripheral and normal skin (P-N) areas. ◇: grad-V (1/20): Gradient in luminance (V) between the inside and the outside of the tumor with the ROI size of 1/20 of the long side of the image. ⋆: #RGB16: The number of colors used in tumor (T) quantized in 16^3 RGB color. ∗: skew-x (205): Skewness of the distribution of the tumor area whose intensity is less than 205 in the major axis. ‡: correlation-90° (1/4): Correlation property of the luminance of the tumor area in the direction of 90 degree from the major axis of the tumor with the unit size of 1/4 of the long side of the image.

very stressed as long as these two are not misclassified for MSL-m.

In the double-shot model, classification performance depends on adjustment of thresholds in both of the AB-C and A-B classifiers. Therefore, in order to acquire appropriate combination of the SE and SP, we conducted grid search in two dimensional space whose axes correspond to the output of each classifier. Note that the thresholds (equivalent to constant term of the linear models) are determined collectively based on the interaction of the two classifiers while coefficients of the linear models except constant term are determined individually.

IV. RESULTS

Table I summarizes the result of stepwise feature selection. The 'Top 10 features' shows the selected features for each classifier when the number of the selected hit 10. The #features shows the number of the selected features at the last step.

Table II shows the summary of classification performances of the single-shot and double-shot models under different number of features. The performance was evaluated under the leave-one-out cross-validation test. The #features denotes the number of features used in each classifier. Note that each of the AB-C and A-B classifier in the double-shot model uses the same number of features as specified by #features.

SE and SP denote the successful classification ratio of MSL-m (A) and the rest (B, C) respectively as mentioned in III-C. We adjusted the classification thresholds to maximize SE×SP.

Error:X→Y means the ratio of X misclassified for Y. Here, error:B→C and error:C→B were not counted since the differentiation of these two is not necessarily important as discussed in III-C. The receiver operating characteristic (ROC) curve is drawn and area under the ROC curve (AUC) is calculated for the single-shot model. AUC is not shown for the double-shot model because the model has two threshold values, making it difficult to define a proper value corresponding to AUC.

V. DISCUSSION

From table I, we can see in the double-shot model that the AB-C is assigned many asymmetry features while the A-B is assigned many color features. On the other hand, the A-BC is assigned both types of features altogether. This indicates that the A-BC needs to conduct both tasks of the AB-C and A-B all at once. Here, since the A-BC has notably smaller number of features (17) compared to the AB-C (33) or A-B (34), we see that it is comparatively difficult to find effective features for the A-BC due to dataset C (NoMSLs) being included in the classification target.

According to max[SE×SP] in table II, the single-shot is better than the double-shot when the selected features are less than 4. However, the double-shot becomes far better than the single-shot with 4 or more features. Actually, the double-shot with only 4 features outperforms the single-shot with 10 features. The highest max[SE×SP] in the single-shot was 0.856 with 17 features while the double-shot with 10 features achieved equivalent or better performance.

As for error ratio in the double-shot, with number of features 4 or more, "error:A→B or C" becomes quite low while error:B→A and error:C→A decrease significantly with the features around 10 or more. In the single-shot, "error:A→B or C" decreases according to increment of features but the other errors do not decrease as much as those in the double-shot.

In most cases in the single-shot model, error:C→A is much larger than error:B→A. This is probably because dataset C (110 images) has less images than dataset B (506 images). As the single-shot tries to distinguish A from B and C by a single classifier, it puts greater emphasis on B over C in feature selection, leading to the poor performance for the

TABLE II
PERFORMANCE OF THE TWO CLASSIFICATION SYSTEMS WITH DIFFERENT NUMBER OF FEATURES

#features	max[SE×SP] (AUC: SE%, SP%) ∗ single-shot (SS)	max[SE×SP] (AUC: SE%, SP%) ∗ double-shot (DS) ⋆	error:A→B or C (%) † SS	error:A→B or C (%) † DS	error:B→A (%) SS	error:B→A (%) DS	error:C→A (%) SS	error:C→A (%) DS
2	0.670 (0.872: 90.5, 74.0)	0.628 (69.0, 90.9)	9.5	31.0	20.9	6.7	49.1	20.0
3	0.683 (0.907: 76.2, 89.6)	0.619 (69.0, 89.6)	23.8	31.0	6.5	6.9	28.2	26.4
4	0.700 (0.913: 78.6, 89.1)	0.783 (97.6, 80.2)	21.4	2.4	5.9	19.8	33.6	20.0
5	0.690 (0.918: 76.2, 90.6)	0.763 (100, 76.3)	23.8	0.0	7.5	23.9	18.2	22.7
6	0.680 (0.916: 83.3, 81.7)	0.813 (95.2, 85.4)	16.7	4.8	14.6	14.0	35.5	17.3
10	0.779 (0.944: 92.9, 83.9)	0.900 (97.6, 92.2)	7.1	2.4	12.8	7.1	30.9	10.9
17	0.856 (0.968: 97.6, 87.7)	0.916 (97.6, 93.8)	2.4	2.4	10.3	6.3	21.8	5.5
33		0.981 (100, 98.1)		0.0		2.0		1.8

∗: Performance was evaluated under leave-one-out test.
⋆: AUC value is not available due to plural thresholds.
†: "Error:A→B or C" means the ratio of A (MSL-m) misclassified for B (MSL-n) or C (NoMSLs).

smaller dataset C. These problems did not happen to the double-shot since it isolates C then B in sequence. Since the classification task is divided into easier subtasks, the stepwise selection yielded more effective features, as indicated by the performance improvement.

However, the double-shot model has poor performance when #features was less than 4. This is probably because at least one classifier had insufficient performance due to lack of efficient features and it caused a severe degradation of the overall performance.

The double-shot model with 10 features achieved an SE of 97.6% and an SP of 92.2%. Although it is difficult to compare results directly since different datasets were used, our result is established on comparatively large number of images and showed superior performance to other state-of-the-art methods reported in [12]. Even though distinguishing melanomas from NoMSLs is considered to be easy for experts, NoMSLs are commonly seen in real world. This achievement should be help of overall melanoma detection performance.

VI. CONCLUSION

In this paper, we presented two melanoma detection models, namely the single-shot and the double-shot, that handle both melanocytic skin lesions (MSLs) and non-melanocytic skin lesions (NoMSLs). Under the leave-one-out cross validation test, the double-shot achieved an SE of 97.6% and an SP of 80.2% with 4 features each. This is better than the single-shot with 10 features. The double-shot divides the classification task into two smaller subtasks and makes it easier to find effective features, leading to superior performance to the single-shot model except the number of features is limited. We are planning on applying the proposed models to our Internet-based melanoma screening system in near future.

ACKNOWLEDGMENT

This research was partially supported by the Ministry of Education, Culture, Science and Technology of Japan (Grantin-Aid for Young Scientists program (B), 23791295, 2011-2012).

REFERENCES

[1] J.M.Olmedo, K.E.Warschaw, J.M.Schmitt, D.L.Swanson, "Correlation of thickness of basal cell carcinoma by optical coherence tomography in vivo and routine histologic findings: a pilot study," Dermatologic Surgery, Vol.33, No.4, pp.421-426, 2007.
[2] K.Nakayama, "Growth and progression of melanoma and non-melanoma skin cancers regulated by ubiquitination," Pigment Cell and Melanoma Research, Vol.23, No.3, pp.338-351, 2010.
[3] A.F.Jerant, J.T.Johnson, C.D.Sheridan, T.J.Caffrey, "Early detection and treatment of skin cancer," American Family Physician, Vol.62, No.2, pp.357-368, 2000.
[4] C.M.Balch, A.C.Buzaid, S.Soong, M.B.Atkins et al., "Final version of the American joint committee on cancer staging system for cutaneous melanoma," Journal of Clinical Oncology, Vol.19, No.16, pp.3635-3648, 2001.
[5] W.Stolz, O.B.Falco, P.Bliek et al., "Color Atlas of Dermatoscopy - 2nd enlarged and completely revised edition," Berlin: Blackwell publishing, 2002.
[6] J.Mayer, "Systematic review of the diagnostic accuracy of dermoscopy in detecting malignant melanoma," The Medical Journal of Australia, Vol.167, No.4, pp.206-210, 1997.
[7] G.Argenziano, H.P.Soyer, S.Chimenti et al., "Dermoscopy of pigmented skin lesions: Results of a consensus meeting via the Internet," Journal of American Academy of Dermatology, Vol.48, No.5, pp.679-693, 2003.
[8] H.Ganster, A.Pinz, R.Rohrer, E.Wilding, M.Binder, H.Kitter, "Automated melanoma recognition," IEEE Transactions on Medical Imaging, Vol.20, No.3, pp.233-239, 2001.
[9] S.Seidenari, G.Pellacani, C.Grana, "Pigment distribution in melanocytic lesion images: a digital parameter to be employed for computer-aided diagnosis," Skin Research and Technology, Vol.11, No.4, pp.236-241, 2005.
[10] M.E.Celebi, H.A.Kingravi, B.Uddin et al., "A methodological approach to the classification of dermoscopy images," Computerized Medical Imaging and Graphics, Vol.31, No.6, pp.362-371, 2007.
[11] H.Iyatomi, H.Oka, M.E.Celebi et al., "An improved Internet-based melanoma screening system with dermatologist-like tumor area extraction algorithm," Computerized Medical Imaging and Graphics, Vol.32, No.7, pp.566-579, 2008.
[12] H.Iyatomi, H.Oka, M.E.Celebi et al., "Computer-based classification of dermoscopy images of melanocytic lesions on acral volar skin," Journal of Investigative Dermatology, Vol.128, pp.20449-2054, 2008.
[13] K.A.Norton, H.Iyatomi, M.E.Celebi et al., "Three-phase general border detection method for dermoscopy images using non-uniform illumination correction," Skin Research and Technology, Vol.18, No.3, pp.290-300, 2012.
[14] H.Iyatomi, K.A.Norton, M.E.Celebi et al., "Classification of melanocytic skin lesions from non-melanocytic lesions," Proc. IEEE EMBS 2010, pp.5407-5410, 2010.
[15] W.Stolz, A.Riemann, A.B.Cognetta, et al., "ABCD rule of dermatoscopy: a new practical method for early recognition of malignant melanoma," European Journal of Dermatology, Vol.4, No.7, pp.521-527, 1994.
[16] B.S.Everitt, G.Dunn, "Applied Multivariate Data Analysis," London: Edward Arnold, pp.219-220, 1991.