Estimation of an optimal spectral band combination to evaluate skin disease treatment efficacy using multi-spectral images

Sylvain Prigent, Didier Zugaj, Xavier Descombes, Philippe Martel, Josiane Zerubia

ESTIMATION OF AN OPTIMAL SPECTRAL BAND COMBINATION TO EVALUATE SKIN DISEASE TREATMENT EFFICACY USING MULTI-SPECTRAL IMAGES

Sylvain Prigent[1], Didier Zugaj[2], Xavier Descombes[1], Philippe Martel[2], Josiane Zerubia[1]

[1]EPI Ariana, INRIA/I3S, 2004 route des Lucioles, BP93, 06902 Sophia Antipolis, Cedex, France
Sylvain.Prigent@inria.fr, Xavier.Descombes@inria.fr, Josiane.Zerubia@inria.fr
[2]Galderma R&D, 2400 Route des Colles, BP87, 06902 Sophia-Antipolis, Cedex, France
Didier.Zugaj@galderma.com, Philippe.Martel@galderma.com

ABSTRACT

Clinical evaluation of skin treatments consists of two steps. First, the degree of the disease is measured clinically on a group of patients by dermatologists. Then, a statistical test is used on obtained set of measures to determine the treatment efficacy. In this paper, a method is proposed to automatically measure the severity of skin hyperpigmentation. After a classification step, an objective function is designed in order to obtain an optimal linear combination of bands defining the severity criterion. Then a hypothesis test is deployed on this combination to quantify treatment efficacy.

Index Terms— hypothesis tests, t-test, multispectral, skin, hyperpigmentation

1. INTRODUCTION

One of the steps to evaluate the efficacy of a therapeutic solution is to perform measurements on a series of patients who received the studied treatment. In parallel another treatment is tested on another group of people or on the same group of patients on another skin area. This second treatment is the reference one for the studied pathology or a placebo. In the following we will use the word 'vehicle' to refer to this treatment. For facial hyper-pigmentation, for each studied treatment, a group of N_e patients receives the treatment on one cheek and the vehicle on the other. To this end patients are selected to have the same hyper-pigmentation severity on the two cheeks. Then severity measurements are taken at different time t along the treatment period. The standard method for assessing effects of each treatment is to clinically scale the disease severity of patients one by one with several dermatologists. This protocol is expensive, long and tedious. Furthermore it is especially prone to the variability of measures that manifests as soon as these are performed by a human being.

To make the process more reliable, we propose an image processing tool that quantifies automatically the disease severity, and then, evaluates the treatment efficacy. To this end, multi-spectral images are used. The process can be splitted in three steps: classification, severity measurement, statistical analysis of the results. To classify skin images of hyper-pigmentation, several algorithms have been proposed in the literature. The benchmark is the $CIEL^*a^*b$ decomposition [1] using color images. The L^* component or the *ITA* index [2] calculated with components L^* and b^* allows to describe the pigmentation. To introduce the spectral information, an algorithm proposed in [3] performs a source separation of the melanin and haemoglobin components from an empirical analysis of their absorption. Methods based on image processing techniques have been proposed in [4, 5] to perform a classification of healthy/pathological areas with multi-spectral images.

Once the healthy and the pathological areas are classified, a criterion has to be defined. This criterion computes the 'darkness' of the skin. So, values of this criterion on both healthy and pathological areas provide the severity measurements. To define this criterion, only the spectral information is considered. It is then characterized by a vector λ such that the measure or darkness for pixel p is a linear combination of spectral bands b:

$$M_p = \sum_{i=1}^{N_b} \lambda_i b_{p,i}, \tag{1}$$

with M_p the value of the criterion at pixel p, N_b the total number of bands in the multi-spectral image and λ_i the i^{st} coefficient of the linear combination $\lambda = \{\lambda_1, ..., \lambda_{N_b}\}$. Since the severity on each patient at every measurement time t is calculated, a hypothesis test is deployed to measure if the severity of the disease evolves significantly. The used test is a *Student test* (*t-test*) [6] which quantifies the deviation of the mean value between two distributions:

$$Z^{t,t_0}(t) = \frac{\bar{X}(t) - \bar{X}(t_0)}{\sqrt{\frac{\sigma^2(t)}{N_e} + \frac{\sigma^2(t_0)}{N_e}}}, \tag{2}$$

where $\bar{X}$ and σ are the mean value and the standard deviation of the two distributions respectively. In our application, a distribution is a set of severity measurements on N_e patients. The *t-test* is performed between a measure at time t and at the baseline at the time t_0. The null hypothesis is: "The mean

value of the distribution does not evolve from t_0 to t". If the probability of false alarm (p-value) of the test is bellow 0.05 the null hypothesis is rejected. And, the lower the p-value, the higher statistically is the deviation between the mean values of the two distributions, and the treatment is proved to be efficient.

The paper is organised as follows. The second section defines the normalization of the data to measure the disease severity. The third section describes the objective function proposed to find an optimum severity criterion per treatment. Then the fourth section shows the obtained results and a comparison with the benchmark method.

2. DATA AND NORMALISATION

For each treatment, the data set of a study is composed of N_e patients. Each patient has two treatments, the active product and the vehicle, in two different areas. Thus there are two multispectral images per patient. After classification of healthy and pathological areas, an average spectrum is computed for each region. Then, four spectra are obtained per patient and per time measurement: the healthy and pathological spectra for the active product and for the vehicle. It comes that two normalizations are needed to compute a severity value from a criterion M. The first one is computed to normalize the criterion measurement between healthy and pathological area, and the second to normalize the measurement of the active treatment by the vehicle. For the healthy/pathological normalization there are two possibilities:

$$d_t^e = \mu_{M_h} - \mu_{M_p} \qquad (3)$$

and

$$d_t^e = \frac{\mu_{M_h} - \mu_{M_p}}{\mu_{M_h}}, \qquad (4)$$

where d_t^e denotes the severity of the disease for the patient e at the measurement time t, and μ_{M_h} and μ_{M_p} are respectively the average measurement of the darkness due to the criterion M on the healthy and the pathological area respectively. The equation (4) seems to be the best choice as it computes a deviation percentage of the pathological area with respect to the healthy one. Nevertheless, the quotient introduces additive noise. And as the two quantities μ_{M_h} and μ_{M_p} are homogeneous the normalization defined in equation (3) is preferred.

A second normalization is needed to compare the active treatment and the vehicle. With the two images taken from the same patient, there is again the two possible normalizations:

$$D_t^e = d_t^{e,A} - d_t^{e,V} \qquad (5)$$

and

$$D_t^e = \frac{d_t^{e,A} - d_t^{e,V}}{d_t^{e,V}}, \qquad (6)$$

where $d_t^{e,A}$ and $d_t^{e,V}$ are the severities measured at the time t on the patient e on the image concerned by the active treatment A, and the vehicle treatment V respectively. For the same reason as above, the chosen normalization is the one of equation (5). Then the final measurement of the severity according to the studied treatment A compared to the reference treatment V is:

$$D_t^e = (\mu_{M_h,A} - \mu_{M_p,A}) - (\mu_{M_h,V} - \mu_{M_p,V}) \qquad (7)$$

It can be noticed that this normalization is linear.

3. THE OBJECTIVE FUNCTION

To determine the severity D_t^e of the patient e at the time t, a criterion M is needed. In this section, we propose to define an objective function to automatically estimate a criterion M which is optimal to describe the treatment effects. If a treatment has an effect, the deviation between the healthy and the pathological areas decreases from the measurement times t to $t + dt$. So, the best criterion will be the one which describes this evolution with the finest scale. This is why, we design an objective function that maximises both the deviation between healthy and pathological areas and the deviation of the severity between time measurements:

$$\mathop{\Delta}_{t-t_0} \left[\frac{1}{N_e} \sum_{e=1}^{N_e} \left[D_t^{e,M} \right] \right], \qquad (8)$$

where Δ is the differential between the initial time t_0 and the time t:

$$\mathop{\Delta}_{t-t_0} (X_t) = X_{t_0} - X_t. \qquad (9)$$

Since the objective is to maximize the deviation mentioned above for any measurement time t with respect to the criterion M, the optimisation problem can be summarized by:

$$\hat{\lambda} = \operatorname*{argmax}_{\lambda} f(\lambda), \qquad (10)$$

where f is the objective function defined by:

$$f(\lambda) = \sum_{t=t_0+dt}^{N_t} \left\{ \mathop{\Delta}_{t-t_0} \left[\frac{1}{N_e} \sum_{e=1}^{N_e} \left[D_t^{e,\lambda} \right] \right] \right\} \qquad (11)$$

where λ is the vector of coefficients defining the criterion M and N_t the number of time measures. To optimize the expression given by equation (10), the geometric simplex method is used [7].

As the final value which quantifies the efficacy of a treatment comes from a *t-test* one can propose to directly minimize the p-value of the *t-test*. Nevertheless, the obtained spectrum will be correlated neither with the disease, non with the treatment effect. That's why we prefer not to do so.

4. EXPERIMENTAL RESULTS

4.1. The data set

The used data are 384 multispectral images taken from 48 patients during 3 months with one measure at the first visit and one every month. Multi-spectral images contain 18 bands in the visible and near infra-red spectra. The group of 48 patients is partitioned in 3 sub-groups of 16 patients. Each sub-group corresponds to a specific tested treatment.

4.2. The treatments

The first treatment is the benchmark for skin pigmentation. We will note it S_t. The two other treatments are from the same product with two different doses. It is the tested product. The two treatments will be called A_{d2} and A_{d3} where d stand for dose and $d2 < d3$. A is written for Active. For the vehicle, we use the active product with a low dose. We will note it A_{d1} with $d1 < d2$. Thus on the 3 groups we measure the treatments S_t vs A_{d1}, A_{d2} vs A_{d1} and A_{d3} vs A_{d1}.

4.3. The baseline

To analyse our method, we use the L^* measure from the $CIEL^*a^*b$. To obtain L^*, a spectral integration algorithm projects the multi-spectral image to the XYZ colours space. Then the transformation from XYZ space to $CIEL^*a^*b$ space allows us to extract the L^* component.

4.4. The results

To avoid some bias due to the classification algorithm, pathological and healthy areas are manually selected on each image. Those areas are manually registered to be in correspondence for each patient along the time sequence. Fig. 2 shows the p-values obtained by the criterion computed with the optimization method. P-values below 0.05 are highlighted with bold. The tested product with a $d1$ dose has an efficacy equivalent to the standard treatment S_t. The null hypothesis of the *t-test* is accepted with a high p-value for all the time sequences. When the dose of the product A increases, it effect should increase and the higher the dose, the longer the effect. The p-values of the *t-test* with the doses A_{d2} and A_{d3} describe well this behaviour. The criterion is adapted to quantify hyper-pigmentation. Fig. 3 summarizes the p-value of the same test but with the Luminance (L^*) as a criterion. This test shows despite that S_t and A_{d1} are equivalent, for higher doses of A, the p-values do not reflect as well as previously the treatment effect. Indeed, for the first measurement, after one month, the effect is not visible; the p-values are above 0.05. Then, along the time sequence, even if the p-values for the treatment A_{d3} follow the expected variations, the p-values variations for A_{d2} are less coherent since they decrease under the 0.05 threshold and then goes back above this threshold.

Thus, the conclusions are that the proposed criterion obtained by optimisation is more sensitive to small variations than the standard criterion L^*. To illustrate this sensitivity, we show in Fig. 1 an example of the evolution of one patient that dermatologists estimate cured at the end of the treatment. As one can see, the measurement follows precisely the disease evolution.

We propose now an interpretation of the criteria. The Luminance is obtained from the component Y, by spectral integration on the multi-spectral images. As a result, L^* in the proposed process is equivalent to use the bands combination Y shown in Fig. 5. Fig. 4 represents the λ coefficients depending on the wavelength obtained by the proposed method. All the obtained criteria have similar shapes but with small variations due to there optimality to each treatment. If we compare the obtained criterion curves with the absorbency curves of melanin and haemoglobin (see Fig. 6) it can be noticed that there is a link between the two. In fact, coefficients of linear combinations are high, in absolute value, when the deviation between haemoglobin and melanin absorbency is high. This observation explains why the proposed method is more efficient to describe the effect of a treatment. When L^* measures only the middle of the spectrum, the objective function allows to select adapted spectral bands.
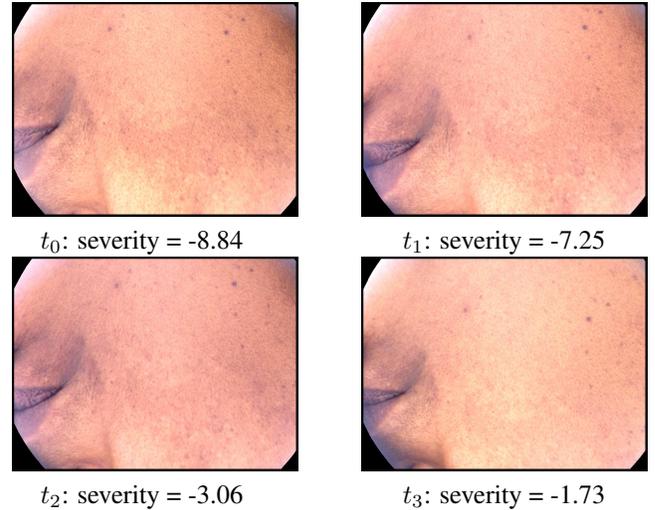

t_0: severity = -8.84 t_1: severity = -7.25

t_2: severity = -3.06 t_3: severity = -1.73

Fig. 1. Illustration of the treatment effect measurement with the proposed method for the treatment A_{d3}.

	$t_1 - t_0$	$t_2 - t_0$	$t_3 - t_0$
S_t vs A_{d1}	$4.027\ 10^{-1}$	$6.545\ 10^{-1}$	$2.867\ 10^{-1}$
A_{d2} vs A_{d1}	$\mathbf{1.158\ 10^{-2}}$	$\mathbf{4.379\ 10^{-3}}$	$\mathbf{3.573\ 10^{-3}}$
A_{d3} vs A_{d1}	$\mathbf{2.240\ 10^{-2}}$	$\mathbf{5.282\ 10^{-3}}$	$\mathbf{5.458\ 10^{-4}}$

Fig. 2. P-values obtained by minimizing the objective function on a spectrum from 400 to 700 nm.

	$t_1 - t_0$	$t_2 - t_0$	$t_3 - t_0$
S_t vs A_{d1}	$9.209\ 10^{-1}$	$5.977\ 10^{-1}$	$9.071 10^{-1}$
A_{d2} vs A_{d1}	$2.577\ 10^{-1}$	$\mathbf{1.478\ 10^{-2}}$	$\mathbf{8.625\ 10^{-2}}$
A_{d3} vs A_{d1}	$1.072\ 10^{-1}$	$\mathbf{4.833\ 10^{-3}}$	$\mathbf{2.996\ 10^{-5}}$

Fig. 3. P-values obtained with the Luminance (L^*) as the darkness measure.
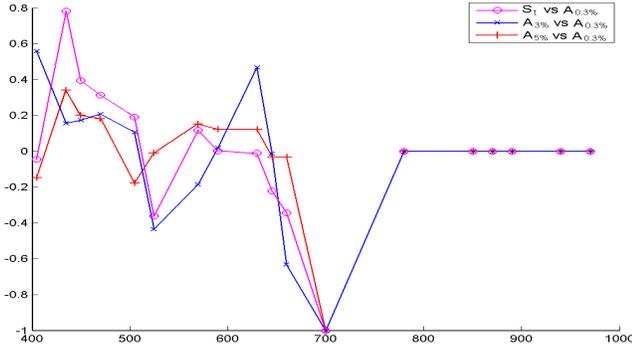
Fig. 4. Coefficients of bands computed by the minimisation of the objective function for every couple of tested treatments.
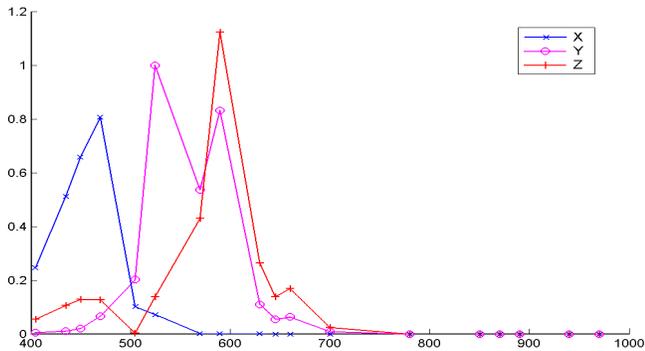

Fig. 5. Coefficient of the linear combination of spectral band to obtain the XYZ decomposition depending on the wavelength.
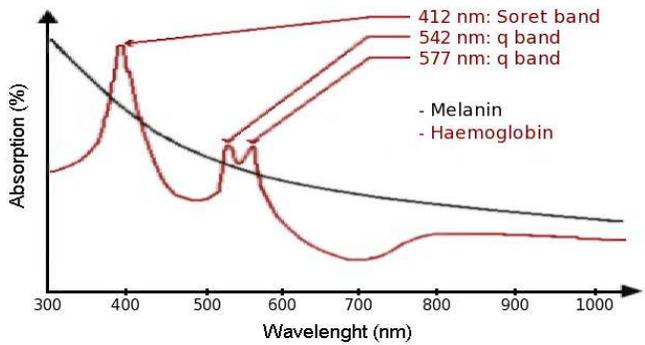

Fig. 6. Absorbency curves of melanin and haemoglobin depending on the wavelength.

5. CONCLUSION

In this paper, we have proposed a method to automatically evaluate the efficacy of treatments with multi-spectral images. The optimization of an objective function allows to find an optimal spectral metric for each tested treatment. We show that such a metric gives a better analysis of the treatments than the standard method. In future work, the spatial information will be studied. In fact, only the spectral information of the average value of a lesion has been exploited. The spatial distribution can give information on the disease and it evolution.

6. ACKNOWLEDGEMENT

7. REFERENCES

[1] I.L. Weatherall and B.D. Coombs, "Skin color measurements in terms of CIELAB color space value," *Journal of Investigative Dermatology*, vol. 99, pp. 468–473, 1992.

[2] G. N. Stamatas, B. Z. Zmudzka, N. Kollias, and J. Z. Beer, "Non-invasive measurements of skin pigmentation in situ.," *Pigment cell res*, vol. 17, pp. 618–626, 2004.

[3] G. N. Stamatas, B. Z. Zmudzka, N. Kollias, and J. Z. Beer, "In vivo measurement of skin erythema and pigmentation: new means of implementation of diffuse reflectance spectroscopy with a commercial instrument," *British Journal of Dermatology*, vol. 159, pp. 683–690, 2008.

[4] S. Prigent, X. Descombes, D. Zugaj, and J. Zerubia, "Spectral analysis and unsupervised SVM classification for skin hyper-pigmentation classification," in *IEEE Workshop on Hyperspectral Image and Signal Processing : Evolution in Remote Sensing (Whispers)*, Reykjavik, Iceland, June 2010.

[5] S. Prigent, X. Descombes, D. Zugaj, P. Martel, and J. Zerubia, "Multi-spectral image analysis for skin pigmentation classification," in *Proc. IEEE International Conference on Image Processing (ICIP)*, Hong-Kong, China, September 2010.

[6] A. M. Mood, F. A. Graybill, and D. C. Boes, *Introduction to the theory of statistics*, McGraw-Hill, 1974.

[7] J.C. Lagarias, J. A. Reeds, M. H. Wright, and P. E. Wright, "Convergence properties of the nelder-mead simplex method in low dimensions," *SIAM Journal of Optimization*, vol. 9, pp. 112–147, 1998.