A METHOD OF DIGITAL DEPARAFFINING BASED ON RAMAN SPECTROSCOPY AND INDEPENDENT COMPONENT ANALYSIS – APPLICATION TO MELANOMA EARLY DIAGNOSIS

C. Gobinet*, A. Tfayli**, O. Piot**, V. Vrabie* and R. Huez*

* CReSTIC, Université de Reims Champagne-Ardenne, Campus du Moulin de la Housse, B.P. 1039, 51687 Reims Cedex 2, France
** Unité MéDIAN, CNRS UMR 6142, Université de Reims Champagne-Ardenne, Faculté de Pharmacie, 51 rue Cognacq Jay, 51096 Reims Cedex, France

{cyril.gobinet, ali.tfayli, olivier.piot, valeriu.vrabie, regis.huez}@univ-reims.fr

Abstract: The paraffin embedding process enables to conserve biopsies for several years. However, the use of paraffinised tissues for spectroscopic investigations remains very restricted. This is due to the intense Raman peaks of paraffin that mask important vibrational information of the tissue in recorded spectra. Paraffinised tissues are usually chemically dewaxed and rehydrated when molecular analysis is required. This is a time consuming procedure and the use of chemical reagents induces alterations in the tissue structure. This paper proposes a method to digitally deparaffinise samples by the association of Raman spectroscopy and Independent Component Analysis (ICA). The dewaxing of the recorded Raman spectra permits the identification of skin spectrum which allows a melanoma early diagnosis.

Introduction

Microscopic and pathological analysis techniques as Raman spectroscopy based methods and histochemical and immunohistochemical techniques, are usually applied to formalin-fixed paraffin-processed tissues previously dewaxed [1-3]. Two main reasons justify the use of paraffin wax embedded tissues. First, tissues need to be stored after their analysis for future studies. Paraffin is a very good tissue preservative. Second, previously cited tissue analysis techniques require thin slices of the tissue in order to transmit light. But sections can be cut exclusively if the sample texture is sufficiently consistent. The paraffin wax embedded process confers excellent cutting and preserving properties on tissues.

Before applying the previously cited methods, tissue samples must be dewaxed and rehydrated. The removal of paraffin wax and rehydration to the aqueous phase is a common practice for the examination of tumoral tissues, including breast cancer [4] and skin cancer [5]. Three main drawbacks are linked with the dewaxing step. First, this process is time and reagents consuming. For instance in [1], the dewaxing agent is xylene and the protocol is quite complicated. Samples are successively plunged in: two baths of xylene for 5 min and 4 min respectively, two baths of Ethanol about 3 min and 2 min respectively, and finally into a bath of Industrial Methylated Spirits 95% for 1 min. To ensure the removal of residual paraffin, samples are cleaned in a bath of xylene during 18 hours. Second, tissues structures can be altered by the high pressure and temperature conditions required by some dewaxing techniques as heat-mediated antigen retrieval (HMAR) [1]. And third, most popular dewaxing methods are not as efficient as lauded. A residual layer of paraffin has been shown to remain after the dewaxing steps in some parts of tissues [1]. The analysis of these tissues will be thus biased.

The development of an efficient dewaxing process is thus essential to overcome these problems. In this article, we propose a digital dewaxing procedure based on the association of Raman spectroscopy based pathological analysis method and a statistical processing technique called Independent Component Analysis (ICA). To the best of our knowledge, our study is the first one that extracts the Raman spectra of a tissue from its non-dewaxed version.

In the first part we will describe the method of digital deparaffining. The second part will show an application of this method to the early diagnosis of melanoma on skin samples. A comparison with the results given by well known Principal Component Analysis (PCA) is also presented. Finally we will summarize and conclude.

Method

The method to digitally deparaffinise biological samples is made up of three main parts. The first one is the recording of tissue Raman spectra. The second step corresponds to the preprocessing of these recorded spectra in order to accommodate their shape for the last step. In this final step, spectra are processed by Independent Component Analysis (ICA) in order to separate parasite signals of paraffin from informative signals of underlying tissues.

Raman spectroscopy is an optical technique designed for the investigation of the molecular structure of tissues. It is based on the interaction between an excitation light and the sample to be analyzed. Light

ISSN: 1727-1983

scattered in an inelastic manner by the tissue is informative about the vibrational and motion states of molecules. Many molecules are Raman active and each possesses its unique Raman spectrum. Raman spectroscopy is an analytical technique which is furthermore non-destructive. Since several years, Raman spectroscopy has been proved to be an efficient tool for investigations in biomedical applications [4-7].

The development of portable Raman spectrometer opens up new horizons for non-invasive diagnosis of diseases. The use of Raman spectroscopy is thus increasing rapidly. However, diagnosis and investigations are first led on paraffin wax embedded biopsies before being applied on *in vivo* tissues.

Raman spectroscopy is very sensitive to paraffin presence due to the intense paraffin characteristic peaks that mask the important vibrational information of the underlying tissue [1]. Furthermore, paraffin wax embedded tissues are generally fixed on a slide whose chemical composition depends on the considered experiment. This slide is often Raman active in a thin spectral band. Usually, CaF_2 or BaF_2 slides are used. Paraffin and the slide corrupt recorded Raman spectra.

Instead of using a non-efficient traditional chemical dewaxing procedure, we rather analyze physical properties of Raman spectra of waxed samples. As usual, Raman spectra are recorded in several points of the investigated tissue. These all spectra form the data matrix X. Each line corresponds to a spectrum recorded in a point, and each column is associated to a wavenumber. The first and obvious feature is the instantaneousness of data recording because the scattered light is collected by CCD detectors. Physical laws governing Raman spectroscopy mechanisms are well known to be linear. Recorded spectra thus result from a weighted sum of spectra of pure species present in the analyzed tissues. This instantaneous and linear model is written as:

$$X = AS^T = \sum_{j=1}^{M} \underline{a}_j \underline{s}_j^T \qquad (1)$$

where S is the pure species spectra or sources matrix and A is the concentration or mixing matrix. The j^{th} column of A, noted $\underline{a}_j$, represents the concentration profile of the j^{th} pure species identified by the spectrum $\underline{s}_j$ (the j^{th} column of S). M denotes the number of sources or distinct species of the model.

The original data matrix X can be decomposed in two subspaces. The first one, denoted by X_n, is called the noise subspace and is composed by uninteresting information, e.g. paraffin and slide spectra. The second one is the signal subspace X_s and is made up of the tissue spectrum that contains useful vibrational information. In a general mathematical form, the subspace decomposition is modeled by:

$$X = X_n + X_s \qquad (2)$$

with $X_n = \sum_{j=\sigma_1}^{\sigma_{M_1}} \underline{a}_j \underline{s}_j^T$, $X_s = \sum_{k=\sigma_{M_1}}^{\sigma_M} \underline{a}_k \underline{s}_k^T$, where σ is a permutation in $\{1,...,M\}$ and M_1 is the number of sources forming the noise subspace. In our case, the noise subspace is composed of the paraffin and the slide influences. In a first approximation, this subspace can be written as:

$$X_n = \underline{a}_{para} \underline{s}_{para}^T + \underline{a}_{slide} \underline{s}_{slide}^T \qquad (3)$$

where $\underline{a}_{para}$, $\underline{s}^T_{para}$, $\underline{a}_{slide}$ and $\underline{s}^T_{slide}$ are the concentration profiles and the spectra of the paraffin and of the slide respectively. In the same way, the signal subspace is made up of the tissue presence and is expressed by:

$$X_s = \underline{a}_{tissue} \underline{s}_{tissue}^T \qquad (4)$$

where $\underline{a}_{tissue}$ and $\underline{s}^T_{tissue}$ are respectively the concentration profile and the spectrum of the tissue to be analyzed.
The aim is now to determine these two subspaces by estimating $\underline{a}_{para}$, $\underline{s}^T_{para}$, $\underline{a}_{slide}$, $\underline{s}^T_{slide}$, $\underline{a}_{tissue}$ and $\underline{s}^T_{tissue}$.

Before suggesting a method that achieves this goal, recorded spectra need to be preprocessed in order to compensate the irrelevant instrumentation and samples variations. In Raman spectroscopy, recorded spectra are usually polluted by a background or baseline signal coming from the fluorescence of the tissue and the response of the spectrometer. We did not incorporate the baseline in the previous subspaces model because the baseline doesn't have a linear behavior from a spectrum to another. However, it is usually modeled by a polynomial function. The polynomial coefficients have been estimated by an algorithm based on the minimization of an asymmetric truncated quadratic cost function and described in [8]. The processed spectra are obtained by subtracting the correspondent baseline from each recorded spectrum.

The alignment of paraffin and slide peaks from a recorded spectrum to another is also an ordinary problem. The resolution of this problem is primordial because noise and signal subspaces are badly estimated if these peaks are not aligned. In this work, an alignment procedure encountered in geophysical signal processing [9] has been adapted. It consists in upsampling spectra in spectral bands where a peak is localized, in computing the shift between a reference spectra peak (commonly the first spectrum recorded on the tissue) and the other spectra peaks, in shifting back the peaks in order to align their maximums, and finally in downsampling spectra.

The final step consists in estimating the noise and signal subspaces knowing only the data matrix X. However, the linearity and instantaneousness of the model (1) is not sufficient to lead to the right estimation of the mixing matrix A and the sources matrix S. Further assumptions on A and S are necessary. The spectrum of paraffin depicted on figure 1(a) is sparse and essentially composed of few thin and energetic peaks. Furthermore, for a given experiment, the slide is chosen to be the less Raman active as possible. Its spectrum shown on figure 1(b) is sparse and composed of a unique peak localized to a small Raman shift and not overlapped with the characteristic peaks of paraffin. Thus, sparsity and non-overlapping of different peaks lead to the statistical

ISSN: 1727-1983

independence between paraffin spectrum and slide spectrum. A final observation is that the tissue spectrum is usually not sparse and not composed of thin and energetic peaks because the tissue is composed of a lot of different chemical compounds, activating almost the entire spectral range. Combining this with the previous remarks, the mutual independence of the underlying pure species spectra is a verified assumption.

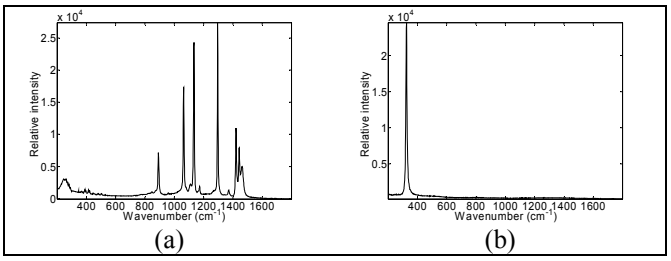

Figure 1: Reference Raman spectrum of (a) paraffin and (b) CaF_2 slide

All conditions are combined to apply ICA techniques [10-11] to the dataset. ICA is a statistical method that blindly estimates the mixing matrix and the sources matrix of the model (1) on the only knowledge of the data matrix X and the mutual statistical independence of columns of S:

$$X = \hat{A}\hat{S}^T = \sum_{j=1}^{M} \underline{\hat{a}}_j \underline{\hat{s}}_j^T \qquad (5)$$

where $\hat{S}$ and $\hat{A}$ are respectively estimators of S and A.

ICA has proved its efficacy to a wide class of applications [12-13]. In this work the Joint Approximate Diagonalization of Eigen-matrices (JADE) algorithm [11] was used to estimate the model components.

Application

Cutaneous melanoma is the most severe form of skin cancers and accounts for three-quarters of all skin cancer deaths. Its incidence and mortality rates are rising worldwide. Clinical diagnosis of malignant melanoma is difficult in numerous cases, particularly due to the difficulty to separate it from atypical benign nevi. Therefore new and efficient non-invasive tools for early diagnosis of melanomas present a crucial interest in clinical practice.

Since few years, several studies have reported the potential of vibrational spectroscopies, infrared absorption and Raman scattering, to characterize and to differentiate cancerous from normal tissues [4-5]. Raman imaging has been often used due to the fact that Raman spectra provide useful information about molecular composition of biological structures.

Recently Tfayli *et al.* [14] have successfully used the FTIR microspectroscopy (a complementary technique of Raman spectroscopy) to discriminate between nevi and melanomas on paraffinised non dewaxed skin sections. However, the discrimination was based on narrow vibrational bands where the paraffin has no contribution.

In this application, we propose first to numerically deparaffinise the Raman spectra by the previous numerical method. Second, we show the possibility to extract discriminant sources specific to malignant and benign tumours, sources that can be employed as molecular descriptors of the type of pathology.

Tissue sections of 10µm thick were cut from paraffin embedded biopsies (Dermatology department of Reims University Hospital) of a malignant melanoma and a benign nevus. Sections were fixed on CaF_2 slides suitable for Raman analysis. Spectral images were collected by a Labram spectrometer (Dilor-Jobin Yvon, Lille, France) in a point by point mode with a 10 µm step. The light source was a titanium-sapphire laser exciting at 785 nm. In each point, the spectrum was recorded at 1305 wavenumbers covering a spectral region from 200 to 1800 cm^{-1} with a resolution of 1.22 cm^{-1}.

Due to the fact that most nevi and melanomas affect the skin epidermis in their first step of development, we were interested in the analysis of this layer only. The analysis of each tissue is thus based on the processing of datasets composed of Raman spectra from the skin epidermis. The malignant melanoma and benign nevus datasets contain respectively 152 and 119 spectra. Two recorded Raman spectra from malignant melanoma and benign nevus respectively are shown on figure 2. As it was argued in the previous section, recorded spectra principally exhibit paraffin and slide vibrational information. Skin presence is weakly visible in a narrow band only.

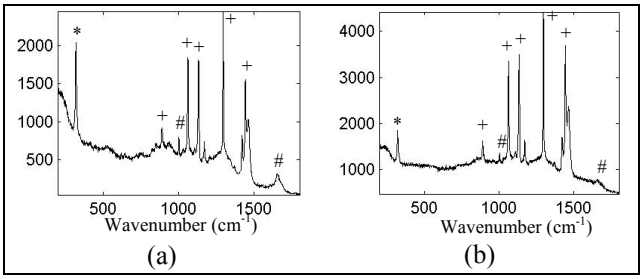

Figure 2: Examples of recorded Raman spectra from (a) melanoma and (b) nevus. Peaks labeled by (*) are associated to the CaF_2 medium and by (+) to paraffin. The visible part of skin spectrum is labeled by (#).

The method described above, composed of the removal of the baseline, the alignment of peaks, and the ICA application, has been applied to the two datasets.

The baseline has been estimated thanks to a five order polynomial function. A higher order polynomial tends to create a very bad estimate of the baseline at the smallest and at the highest wavenumbers. An example of the baseline estimation is given on figure 3(a) for a nevus spectrum. The corrected spectrum resulting from the subtraction of the estimated baseline from the recorded spectrum is shown on figure 3(b). As we can see, this method is very efficient and lead to a very well cancellation of the baseline.

ISSN: 1727-1983

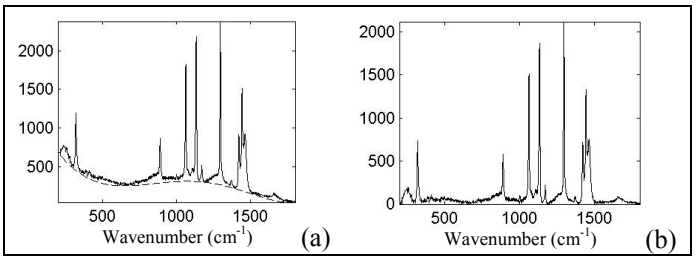

Figure 3: Example of (a) the estimation of the baseline by a five order polynomial function and (b) the removal of this baseline from a recorded spectrum of a nevus.

The peaks alignment procedure is applied for each baseline corrected spectrum for peaks localized at 322, 888, 1061, 1132, 1170, 1294, 1419 and 1442 cm^{-1}. An example illustrates this technique on figure 4. The reference peak (solid line) and the peak to be processed (dashed line) are represented on figure 4(a). Figure 4(b) zooms in the maxima of these peaks. The resulting aligned peak is shown on figure 4(c) and its zoomed version on figure 4(d). We can notice that even if a peak is far from the reference peak, the alignment procedure is very efficient. By "far from the reference peak", it is meant that the maxima of the peak and of the reference are localized to different wavenumbers, and that their widths are different.

The estimation of underlying sources can now be achieved by ICA. A difficulty is the choice of the number of sources model. As skin signal is weak compared to paraffin and slide signals, usual ways for determining the number of sources are useless. The number of independent components has been selected by visual inspection of results. Three or four sources models (M= 3 or 4 in eq. (5)) let the skin spectrum mixed with the paraffin spectrum. A five components model leads to a well decomposition of datasets. The estimated sources are depicted on figure 5 for a benign nevus. Three sources are modeling the spectrum of paraffin. A source is attributed to the CaF_2 slide. And finally, an estimated spectrum is characteristic of skin. A first conclusion is that the paraffin spectrum has not a linear behavior in function of the wavenumbers. A spectral band reacts independently from the others with the underlying tissue. Nevertheless, the interesting information is the skin spectrum, so the decomposition in several sources of paraffin spectrum does not handicap the interpretation of results because paraffin is considered as a polluting component in this application. However, this is an important result concerning the non dewaxed tissues: the paraffin must be modelized by more than one source.

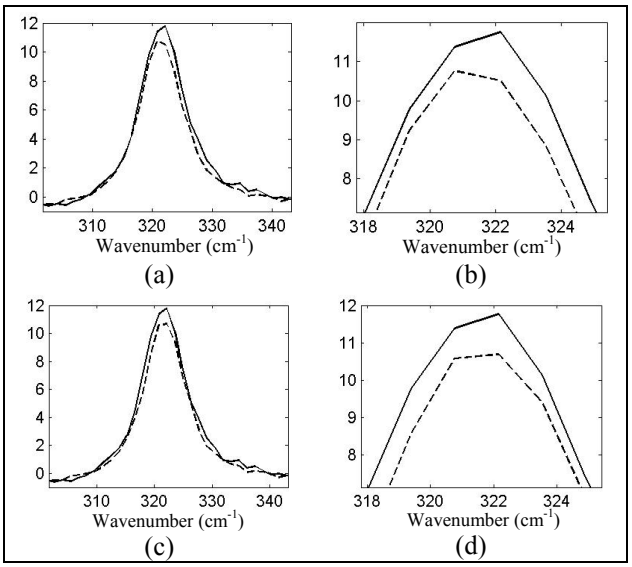

Figure 4: Illustration of the peak alignment process. (a) Recorded reference spectrum (solid line) and spectrum to be aligned (dashed line). (b) Zoom in their maxima. (c) Aligned peak and the reference peak. (d) Zoom in their maxima.

The efficacy of numerical deparaffining can be revealed by the examination of results thanks to a subspaces representation. Figure 6.1. depicts a recorded spectrum of a nevus. Thanks to ICA, it is decomposable in two subspaces: The first one is the noise subspace (figure 6.2) composed by uninteresting information, e.g. paraffin and CaF2 spectra, and the second one (figure 6.3) is the signal subspace composed of skin interesting information. Note that this last one is just a scaled version of the spectrum associated to skin as shown in figure 5(e). We can notice that the signal subspace is not very energetic compared to the noise subspace. Another proof of the efficiency of ICA is the total positivity of estimated concentration profiles, contrary to those estimated by Principal Component Analysis (PCA) as it will be seen later.

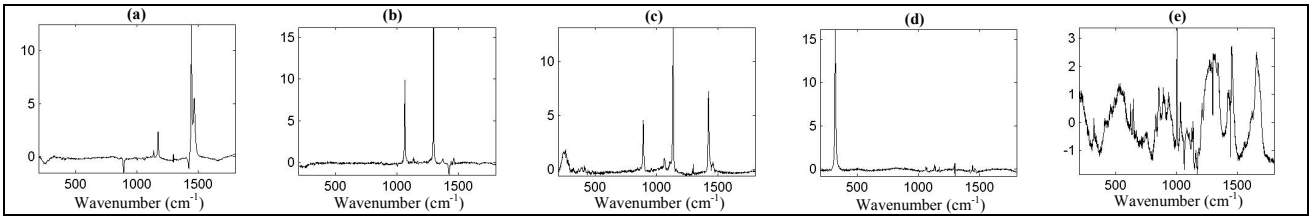

Figure 5: Independent components estimated by an M = 5 sources model on a benign nevus: (a), (b) and (c) independent components associated to paraffin. (d) independent component associated to CaF_2. (e) independent component associated to skin.

ISSN: 1727-1983

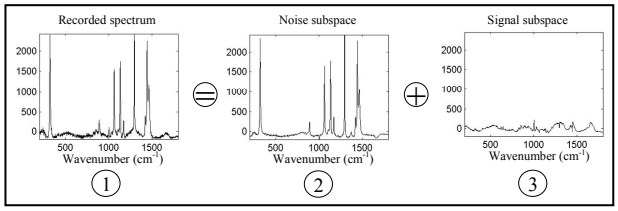

Figure 6: Decomposition of initial data space into a noise subspace and a signal subspace. (*1*) A recorded spectrum lying in the data space. (*2*) Its noise part lying in the noise subspace. (*3*) Its signal part lying in the signal subspace

The previous described method has also been applied to a melanoma dataset. Estimated underlying sources are the same as for the nevus, except the skin spectrum that shows different vibrational information in function of the kind of analyzed skin sample. A comparison of spectra associated to skin estimated for a malignant melanoma and a benign nevus can be done, as shown on figure 7.

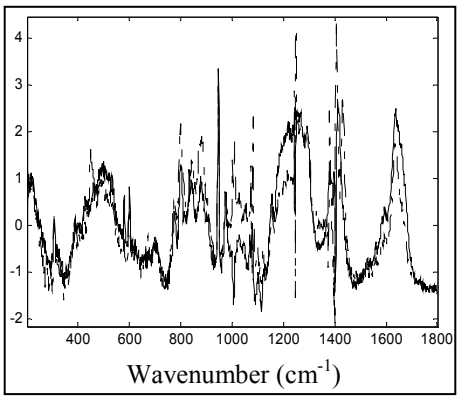

Figure 7: Skin spectra estimated on a benign nevus (*solid line*) and on a malignant melanoma (*dashed-dotted line*)

The sources obtained from ICA show visible differences between nevi and melanomas. Such differences are visualized with the changing intensity ratio of the Fermi doublet bands on 850 and 830 cm^{-1}. For melanomas it is around 2.5 while for the nevi it is only 1.6. Such changes could inform us about the state of the phenylic cycle in the tyrosine residue and the type of resulting molecular bands (intra- or inter-).

Secondary structure variations are marked by a predominance, in the melanoma source, of the α helix vibrations (1650cm^{-1}) in the amide I band. Similar information can be obtained from the high intensity of the band on 934 cm^{-1} characterizing the C-C stretch in the α helix.

On the other hand, the nevi source represents a shoulder band at 1670 cm^{-1} revealing a more important contribution of the β sheet conformation.

The differences in the secondary structure can be quantified by the decomposition of the amide I band by creating spectral models with Gaussian-Loretzian functions.

The same information can be obtained from the changes of the amide III band, and from the intensity of the band on 901 cm^{-1}.

Other differences can be also observed, for example with the band on 1620 cm^{-1} attributed to the DNA, or with the band on 480 cm^{-1} that could be attributed to aromatic amino acids.

Other changes in the amino acid residues content of the protein can be detected with phenylalanine bands on 620, 1003, 1033 and 1610 cm^{-1}, and tyrosine at 640 and 1610 cm^{-1}.

To finish, we want to compare ICA results with those obtained by processing data with Principal Component Analysis (PCA) because PCA is commonly applied to biological and biophysical datasets [4, 15]. PCA is searching for statistically decorrelated sources that respect the linear model (1). The decorrelation is only a second order independence, so estimated spectra may be a linear combination of pure species spectra. It can be seen on figure 8 where the two first principal components are depicted. Even if the study of the principal components may lead to discrimination between tissues, pure species spectra are not well identified. Paraffin and CaF_2 spectra cannot be exactly subtracting. It is shown that the first estimated spectrum at the left of the figure has it peaks well oriented. The second one at the right of the figure exhibits peaks oriented to opposite directions. This is physically unrealistic. Moreover, concentration profiles of these components exhibit negative and positive values. To feat to reality, only totally positive concentration profiles are admitted. Furthermore, the spectra estimated by PCA are still linear combinations of pure species spectra as can be observed in figure 8 where influence of CaF_2 and paraffin are mixed. Thus the decorrelation assumption is not sufficient to lead to an accurate estimation of sources, contrary to the independence hypothesis.

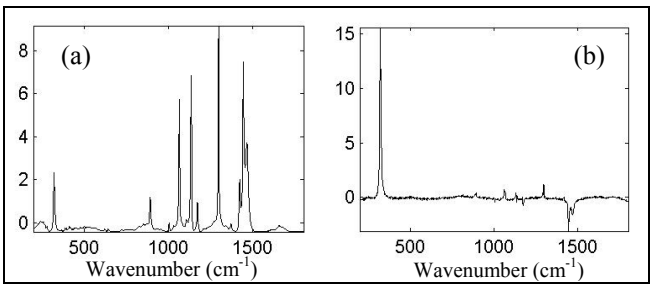

Figure 8: The two first principal components estimated on a benign nevus

Conclusion

The association of Raman spectroscopy and ICA leads to a useful digital deparaffining technique. This

ISSN: 1727-1983

method is based on the decomposition of the recorded spectra in a noise part and a signal part. When applied to skin datasets, spectral differences between nevus and melanoma skin spectra lead to a discrimination of these two kinds of tissue. Consequently, this method allows a melanoma early diagnosis. Furthermore, an important result is that paraffin spectrum cannot be modeled by a unique source. Three sources are necessary to describe the independent variations of characteristic spectral bands of paraffin. However, a limitation of the presented method is the manual choice of the number of underlying sources.

References

[1] Ó FAOLÁIN E., HUNTER M., BYRNE J., KELEHAN P., LAMBKIN H., BYRNE H., and LYNG F. (2005): 'Raman Spectroscopic Evaluation of Efficacy of Current Paraffin Wax Section Dewaxing Agents', *Journal of Histochemistry & Cytochemistry*, **53**, pp. 121-129

[2] LOWRY A., WILCOX D., MASSON, E. A., and WILLIAMS P. E. (1997): 'Immunohistochemical Methods for Semiquantitative Analysis of Collagen Content in Human Peripherical Nerve', *Journal of Anatomy*, **191**, pp. 367-374

[3] ENDL E., KAUSCH I., BAACK M., KNIPPERS R., GERDES J., AND SCHOLZEN T. (2001): 'The Expression of Ki-67, MCM3, and p27 Defines Distinct Subsets of Proliferating, Resting, and Differentiated Cells', *Journal of Pathology*, **195**, pp. 457-462

[4] HAKA A., SHAFER-PELTIER K., FITZMAURICE M., CROWE J., DASARI R., AND FELD M. (2002): 'Identifying Microcalcifications in Benign and Malignant Breast Lesions by Probing Differences in Their Chemical Composition Using Raman Spectroscopy', *Cancer Research*, **62**, pp. 5375-5380

[5] GNIADECKA M., WULF H., MORTENSEN N., NIELSEN O., AND CHRISTENSEN D. (1997): 'Diagnosis of Basal Cell Carcinoma by Raman Spectra', *Journal of Raman Spectroscopy*, **28**, pp. 125-129

[6] BUSCHMAN H., MOTZ J., DEINUM G., RÖMER T., FITZMAURICE M., KRAMER J., VAN DER LAARSE A., BRUSCHKE A., AND FELD M. (2001): 'Diagnosis of Human Coronary Atherosclerosis by Morphology-based Raman spectroscopy', *Cardiovascular Pathology*, **10**, pp. 59-68

[7] CHOO-SMITH L.-P., EDWARDS H., ENDTZ H., KROS J., HEULE F., BARR H., ROBINSON J., BRUINING H., AND PUPPELS G. (2002): 'Medical Applications of Raman Spectroscopy: From Proof of Principle to Clinical Implementation', *Biopolymers: Biospectroscopy*, **67**, pp. 1-9

[8] MAZET V., CARTERET C., BRIE D., IDIER J. AND HUMBERT B. (2005): 'Background Removal from Spectra by Designing and Minimising a Non-quadratic Cost Function', *Chemometrics and Intelligent Laboratory Systems*, **76**, pp. 121-133

[9] VRABIE V., LE BIHAN N., AND MARS J. (2002): '3D-SVD and Partial ICA for 3D Array Sensors', Proc. Of SEG'2002 - 72nd International Conference of the Society of Exploration Geophysicists. Salk Lake City, USA, 2002

[10] COMON P. (1994): 'Independent Component Analysis: A New Concept?', *Signal Processing*, **36**, pp. 287-314

[11] CARDOSO J.-F., AND SOULOUMIAC A. (1993): 'Blind Beamforming for Non-Gaussian Signals', *IEE Proceedings F*, **140**, pp. 362-370

[12] DE LATHAUWER L., DE MOOR B., AND VANDERWALLE J. (2000): 'Fetal Electrocardiogram Extraction by Blind Source Subspace Separation', *IEEE Transactions on Biomedical Engineering*, **47**, pp. 567-572

[13] ZIEHE A., MÜLLER K.-R., NOLTE G., MACKERT B.-M., AND CURIO G. (2000): 'Artifact Reduction in Magnetoneurography Based on Time-Delayed Second-Order Correlations', *IEEE Transactions on Biomedical Engineering*, **47**, pp. 75-87

[14] TFAYLI A., PIOT O., DURLACH A., AND MANFAIT M. (2005): 'Discriminating Nevus and Melanoma on Paraffin Embedded Skin Biopsies using FTIR Microspectroscopy', *BBA General Subjects*, accepted manuscript

[15] SIGURDSSON S., PHILIPSEN P., HANSON L., LARSEN J., GNIADECKA M., AND WULF H. (2004): 'Detection of Skin Cancer by Classification of Raman Spectra', *IEEE Transactions on Biomedical Engineering*, **51**, pp. 1784-1793

ISSN: 1727-1983