Article

Applied Spectroscopy

Applied Spectroscopy
2019, Vol. 73(8) 893–901

DOI: 10.1177/0003702819839098
journals.sagepub.com/home/asp
SAGE

An Algorithm for the Removal of Cosmic Ray Artifacts in Spectral Data Sets

Sinead J. Barton[1] and Bryan M. Hennelly[1,2]

Abstract

Cosmic ray artifacts may be present in all photo-electric readout systems. In spectroscopy, they present as random unidirectional sharp spikes that distort spectra and may have an affect on post-processing, possibly affecting the results of multivariate statistical classification. A number of methods have previously been proposed to remove cosmic ray artifacts from spectra but the goal of removing the artifacts while making no other change to the underlying spectrum is challenging. One of the most successful and commonly applied methods for the removal of comic ray artifacts involves the capture of two sequential spectra that are compared in order to identify spikes. The disadvantage of this approach is that at least two recordings are necessary, which may be problematic for dynamically changing spectra, and which can reduce the signal-to-noise (S/N) ratio when compared with a single recording of equivalent duration due to the inclusion of two instances of read noise. In this paper, a cosmic ray artefact removal algorithm is proposed that works in a similar way to the double acquisition method but requires only a single capture, so long as a data set of similar spectra is available. The method employs normalized covariance in order to identify a similar spectrum in the data set, from which a direct comparison reveals the presence of cosmic ray artifacts, which are then replaced with the corresponding values from the matching spectrum. The advantage of the proposed method over the double acquisition method is investigated in the context of the S/N ratio and is applied to various data sets of Raman spectra recorded from biological cells.

Keywords

Raman spectroscopy, cosmic rays, noise reduction, processing spectra

Date received: 16 November 2018; accepted: 29 January 2019

Introduction

Cosmic ray artefact (CRA) contamination occurs frequently when recording spectra using any photo-electric device, such as a charge-coupled device (CCD). These intermittent events are caused by high-energy particles interacting with the detector,[1] the effect of which is to release large numbers of electrons that are indistinguishable from photoelectrons. Cosmic ray artifacts are randomly distributed in time and intensity and are generally localized to a small number of adjacent pixels in an array detector, although they may, in some cases, have a broader width.[2] Cosmic ray artifacts can be especially prominent when the spectral irradiance is weak, such as for the case of Raman spectra recorded from biological samples, which necessitates a detector that is sensitive to low photon counts and the utilization of long camera integration times.

The distortion of spectra by the presence of CRAs can pose problems for various applications that involve the identification of specific peaks. Cosmic ray artifacts can also impact on the results of post-processing algorithms such as principle component analysis (PCA), due to biasing of the loading vectors towards large outliers, which in turn leads to the misclassification of spectra.[3] The misclassification of spectra can be of critical importance, particularly in the growing area of chemometrics.[4,5]

A number of methods have previously been proposed for the detection and replacement of CRA contaminated pixels. Following from the classification system proposed by Li et al.,[6] these methods fall into four categories. The first category is based on single capture methods, which can significantly impact the underlying spectrum if the CRAs are of a similar width to spectral features.[7–13] The second category provides superior performance in this regard but requires multiple successive captures from the sample.[14–21]

[1]Department of Electronic Engineering, Maynooth University, Kildare, Ireland
[2]Department of Computer Science, Maynooth University, Kildare, Ireland

Corresponding author:
Bryan M. Hennelly, Department of Electronic Engineering, Maynooth University, Maynooth, Kildare, Ireland.
Email: bryanh@cs.nuim.ie

The third category relies on optimized hardware that is resistant to the detection of CRAs.[22] Finally, the fourth category proposes to remove the CRA noise from Raman spectral images.[3,6,23–25] A more detailed description of the these four different types of methods is provided in the Background section.

In this paper, a novel CRA removal algorithm is proposed that combines aspects from the first two categories and has the advantages of both; the method requires only a single capture but works on the same principle as the double acquisition method and provides comparable results, i.e., it removes only cosmic rays and makes no other changes to the spectrum. The method requires the availability of a data set of spectra that can be used for comparison, the most similar of which is identified using normalized covariance. The spectrum of interest is then directly compared with the matching spectrum and differences exceeding a specified threshold are identified as cosmic rays. The contaminated pixels are replaced with the corresponding spectral value from the matching spectrum. The optimal value of the threshold is automatically estimated based on the standard deviation of the noise in the spectrum, which is indicative of the level of noise present in the spectrum. The algorithm can be applied to an entire data set of recorded spectra without intervention from the user.

In addition to being a single acquisition method, the proposed algorithm has a second advantage over the double acquisition method, in that it may offer a significant improvement in the signal-to-noise (S/N) ratio of the denoised spectrum under certain conditions, due to the reduced instances of camera read noise that are included, which is discussed in more detail in the Noise in a Spectrum: Single Versus Double Acquisition section. The requirement for an available data set of spectra is naturally met for a large number of applications that involve the repeated capture of data such as Raman based chemometrics for the detection of bladder cancer,[26] cervical neoplasia,[27] and breast cancer detection.[28] Applications such as these require repeated measurement from cell or tissue samples and, therefore, a data set of related spectra will often be readily available.

Background

Li et al. propose four categories for all CRA removal methods. The first of these comprises single scan methods that rely on the assumption that CRAs will have an appreciably narrower width than the expected peaks in the spectrum. This requires that the spectral resolution of the system is less than the width of the spectral peaks, which may not always be the case and depends on the properties of both the source laser and spectrograph in the recording system as well as the chemical composition of the sample under investigation. Methods in this category include the "missing point polynomial filter",[10,11] the wavelet transform method,[7,13] filtering based on fuzzy logic,[8] weighted moving filters,[9] and median and low pass filtering.[12] In many cases, the methods in this category are unsuitable because they are either insensitive to CRAs that have comparable width to the features of the underlying spectrum or they rely on empirically chosen thresholds that may vary between data sets. As a result, in some cases the denoised spectra must be subjected to robust error checking and this can limit the inclusion of these algorithms in fully automated applications.

The second category of methods for the removal of CRAs is based on the low probability of CRAs contaminating the same pixel in sequential spectra. The algorithms in this category include the upper bound spectrum (UBS) method and its improved variations,[14,15,16,18] based on whether there are sustained changes in the spectral profile,[19–21] and multiple acquisition methods used by manufacturers of commercial optical spectroscopy systems such as Renishaw and Horiba as described in the Noise in a Spectrum: Single Versus Double Acquisition section.[17]

Optimization of optical systems in order to avoid detection of CRAs, such as image curvature correction,[22] is a third option. In this case, CRAs are detected by comparing spectra recorded along different rows of pixels on the detector. Aberration caused by the imaging system may necessitate numerical correction before comparison.

The fourth category is based on a mapping technique[3,23–25] and requires a map of spatially adjacent spectra. A nearest neighbor comparison (NNC) is performed and the most closely correlated spectrum is selected. An offset is selected based on the expected noise and if the intensity value of a spectral component in the original spectrum differs from the corresponding value in the offset spectrum by a value exceeding said offset then the lower value is taken. The algorithm presented in this paper is similar to this approach, except that comparison is performed across an entire data set of spectra rather than over a set of spatially adjacent neighbors.

Although all these methods have been shown to be effective CRA removal methods, in some cases they are computationally intensive or rely on expensive equipment, which may not be feasible. The double acquisition method[17] is arguably the most commonly used approach due to its simplicity and accuracy. The proposed algorithm aims to emulate the robust nature of this method while providing the advantages of single acquisition and improving the resulting S/N ratio of the denoised spectrum.

Noise in a Spectrum: Single Versus Double Acquisition

Noise is any unwanted perturbation in the signal of interest; in the case of Raman spectroscopy, this is considered to be any extraneous electrons that accumulate in the detector. In general, there are four main sources of noise: shot noise; dark current; read noise; and the main focus of this paper,

CRAs. The first two sources of noise are time-dependent, while read noise is time-independent. Cosmic rays occur randomly in both space and time and so this noise source may also be described as time-dependent. Shot noise is the result of inconsistent flux incident on the detector pixels. This discrepancy over time causes jitter in the signal and is governed by a Poisson distribution meaning that the standard deviation of the shot noise is related to the square root of the spectral intensity. The impact of shot noise can be minimized by gathering large numbers of photons such that the inconsistencies become insignificant compared to the collected signal. For weak irradiances, this requires long exposure times.

Dark current, which comprises thermally generated electrons within the charge-coupled device (CCD) detectors semi-conductor pixels, is also modelled by a Poisson distribution. Dark current can be reduced by cooling the detector and recording for short acquisition times. Read noise is introduced through the electronics in the detector that are used to extract the electrons from the pixel wells and digitize the signal. Read noise is inherent to all signal acquisitions and is considered to be the ultimate limiting factor in single photon detection. It is dependent on readout rates and the quantization levels of the analogue to digital converter. While this noise can be minimized by selecting low readout rates or by using an electron-multiplying CCD or modern scientific complementary metal–oxide–semiconductor (sCMOS) detectors, it cannot be fully eradicated.

Multiple acquisitions, whereby a number of spectra are averaged together for the purpose of CRA removal, can have a negative impact on the S/N ratio of the resulting denoised spectrum. Shot noise and dark current noise are both modelled by time-dependent Poisson distributions. Therefore, if only these two noise sources are considered, a spectrum collected with a 5 s acquisition time will have the same S/N ratio to two 2.5 s spectra collected under the same conditions and averaged together. However, read noise is time-independent and will be included in each individual recorded spectrum and, therefore, averaging a number of acquisitions together will introduce multiple instances of read noise. The S/N ratio in a single sample of the spectrum is defined as follows:[29]

$$SNR = \frac{iq(\lambda)tp_i}{\sqrt{[iq(\lambda)p_i + cp_{dc}]t + p_r n_r}} \quad (1)$$

where i represents the spectral irradiance that is incident on each individual pixel in a column of p_i pixels, which depends on the spatial distribution of the light arriving at the spectrograph slit; $q(\lambda)$ is the quantum efficiency of a pixel for the incident wavelength λ and t is the total camera integration time; c is the mean rate of dark current production in electrons per pixel per second; p_{dc} is the number of pixels contributing dark current noise and p_r is the number of pixels contributing read noise to the spectral component. In full vertical bin (FVB) mode, p_{dc} will be the same as the number of pixels in the full column and $p_r = 1$ as there is only one instance of read noise. Some cameras support "crop mode", whereby a FVB can be applied over a reduced number of rows. In this case, it is possible to match the values of p_i and p_{dc} such that dark current is amassed only from pixels that detect photons. Finally, in "image mode" each pixel is read out independently, each with its own instance of read noise. In this case, each row consists of a single spectrum with $p_i = p_{dc} = p_r = 1$. A more detailed discussion on noise contributions for different camera modes is given in Barton and Hennelly.[30]

Using Eq. 1, it is possible to compare the S/N ratio of a single acquisition of time T to N acquisitions, each of time T / N duration, which are subsequently averaged. Assuming the camera mode is consistent, both cases will result in a spectrum that has the same spectral intensity (i.e., the numerator in Eq. 1 will be $iq(\lambda)p_i T$ for both cases). Similarly, the dark current contribution will also be the same. However, the read noise contribution will differ for both cases; for the single acquisition $p_r = 1$ and for the multi-acquisition $p_r = N$. Therefore, it can be expected that the multi-acquisition will have a reduced S/N ratio when compared with a single acquisition of equivalent duration. The difference between these two S/N ratios will be determined by the values of i, c, n_r, and N.

Vendors of commercial systems and cameras often provide their own software such as Horiba SynerJY, Andor Solis, and Renishaw WiRE. The first two of these favor sequential scanning methods to remove CRAs and the third uses a median filtering approach. While these approaches are robust, all of them require multiple captures, which may not be possible for some applications. An approach that combines aspects from sequential scan and NNC methods would be a useful alternative in applications that are time-sensitive and where read noise is a significant contributor to the noise levels.

Proposed Algorithm

The first step in the proposed algorithm is to assign a best matching pair to each spectrum in a given data set, thereby removing the need to record multiple spectra. These pairs of matching spectra are then denoised in a similar manner to that of the commonly used double acquisition method, by identifying corresponding samples in the spectrum for which there exists a difference in intensity that is greater than a threshold that relates to the expected noise level. The final step in the algorithm is to apply a smaller threshold to the immediate neighbors of a sample that has been contaminated with a cosmic ray in order to ensure that even broad CRAs are effectively removed from the spectrum.

Step 1. In order to pair spectra together, an approach similar to NNC[23] is employed, which identifies spectra in a given data set that share a high normalized covariance. The normalized covariance is calculated as follows:

$$C_{nm} = \frac{(S_n \cdot S_m)^2}{(S_n \cdot S_n)(S_m \cdot S_m)} \tag{2}$$

where C_{nm} denotes the normalized covariance of spectra n and m in the data set and "·" represents the dot product. For each spectrum in the data set, i.e., $n = \{0, 1, 2 \ldots N-1\}$, the value of C_{nm} is calculated for all values of $m = 0, 1, 2 \ldots N-1$ where $m \neq n$. For a given spectrum S_n, the spectrum S_m that corresponds to the maximum value of C_{nm} is taken to be the most similar and is paired with S_n for the next stage of the algorithm. In this way, each spectrum in the data set, S_n, is given a pair denoted by $S_{n'}$.

Step 2. As previously discussed in the Noise in a Spectrum: Single Versus Double Acquisition section, a priori knowledge can be used to calculate the standard deviation of the noise in a spectrum; CRAs are then identified as spikes that exceed some threshold that is proportional to this value. A similar approach has been proposed in the double acquisition method.[17] However, this method requires knowledge of the specifications of the spectrometer as well as the expected irradiance, which may not be available. For this reason, we propose a method to estimate the standard deviation of the noise in a given spectrum, n, without any *a priori* knowledge of the recording system, as defined by Eq. 3.

$$\sigma_n = \frac{1}{N}\sqrt{\sum_{k=1}^{M} [S_n(k) - \overline{S_n(k)}]^2} \tag{3}$$

where k is the *kth* sample of the spectrum, the value of k ranges from 1 to M, and $\overline{S_n(k)}$ is the Savitzky–Golay smoothed version of the raw spectrum. If the intensity of the residuals resulting from $S_n(k) - S_{n'}(k)$, exceeds the threshold given by $5\sigma_n$ the pixel is deemed to be corrupted and is replaced with $S_{n'}(k)$. This process is formally defined in Eq. 4 and is repeated for all values of k from 1 to M.

$$S_n(k) = \begin{cases} S_n(k) & \text{if } \mathrm{S_n(k)} - S_{n'}(\mathrm{k}) < 5\sigma_\mathrm{n} \\ S_{n'}(k) & \text{if } \mathrm{S_n(k)} - S_{n'}(\mathrm{k}) > 5\sigma_\mathrm{n} \end{cases} \tag{4}$$

The $5\sigma_n$ threshold ensures that >99% of the noise inherent in the recorded signal, i.e., shot noise, dark current, and read noise will fall within this boundary. The likelihood of a CRA being detected where there is none is <1%. We note that the algorithm described above is similar to the method proposed in Takeuchi and Harada,[17] whereby two spectra are recorded in succession and averaged. Corresponding samples that have a disparity greater than the defined threshold are not averaged and the lesser sample value is taken.

Step 3. It is possible to further amend the algorithm described above in order to deal with the case in which a CRA has a larger width than a single pixel and extends into neighboring pixels although possibly falling under the specified threshold. A reduced threshold can be applied to the pixels immediately around a detected CRA; this process is formally defined in Eq. 5 and is repeated for each value of k corresponding to the sample location of a detected CRA in Step 2.

$$S_n(k \pm 1) = \begin{cases} S_n(k \pm 1) & \text{if } \mathrm{S_n(k \pm 1)} - S_{n'}(\mathrm{k \pm 1}) < 2\sigma_\mathrm{n} \\ S_{n'}(k \pm 1) & \text{if } \mathrm{S_n(k \pm 1)} - S_{n'}(\mathrm{k \pm 1}) > 2\sigma_\mathrm{n} \end{cases} \tag{5}$$

This addition improves the overall sensitivity of the algorithm to include broader CRAs.

In the case of biological spectra, varying baselines and sample heterogeneity can produce significant inconsistency across the spectra in the data set, which can reduce the capability of the proposed method to find an accurate match within the data set for a given spectrum in Step 1. In this case, it is recommended to perform a pre-processing step in the form of a background subtraction algorithm[31–33] on the data set in order to reduce variability and ensure a high correlation between matched spectra. The background subtraction algorithm used in this paper is described in Afseth and Kohler[31] and is based on first calculating the mean spectrum for the whole data set, followed by least squares fitting of an N order polynomial as well as the mean spectrum to each individual spectrum. This step can easily be reversed following the CRA removal algorithm by reintroducing the subtracted baseline back to each respective spectrum, if desired.

In systems and applications for which there is an increased possibility of having CRA contamination at the same sample point in multiple spectra, it may be required to apply median filtering in advance of matching pairs of spectra. Any system where multiple spectra are acquired simultaneously, e.g., a line illumination system, is prone to a single CRA contaminating a number of spectra in the same pixel region. Large data sets that are obtained using long acquisition times are also susceptible to this as they will contain a large number of CRAs and, therefore, the likelihood of a CRA appearing at the same sample point across multiple spectra in the data set increases. Due to the intensity of these spikes, it is likely that Step 1 of the algorithm will match these spectra together due to their high covariance. In order to avoid this, a median filter can be applied to the data set and the normalized covariance in Step 1 may be calculated based on this filtered data set.

A flow chart of the overall algorithm including these preprocessing steps is illustrated in Fig. 1. In the sections that follow, the proposed algorithm is applied to data sets of

Raman spectra and the performance is compared to that of the double acquisition method.

Materials and Methods

Materials

A polymer reference material, acquired from Ibidi GmbH,[34] was chosen as the first sample for investigation due to its thermal stability, resistance to photo-bleaching, and strong reliable signal, which reduces the overall experimental variability. The consistency of this sample and its insignificant baseline ensures an accurate assessment of the proposed CRA algorithm in terms of the S/N ratio. The benefits of a single instance of read noise in terms of S/N ratio will be more significant for weak spectral irradiances such as for the case of a Raman spectrum recorded from a biological sample. Ideally, a biological sample would have been used to demonstrate the improvement in S/N ratio afforded by the proposed algorithm when compared with the double acquisition method. However, photo-bleaching and the heterogeneity of biological samples may complicate an accurate measurement of S/N ratio. It was, therefore, decided to use the polymer sample for the evaluation of the proposed algorithm in terms of S/N ratio and to reduce the recorded irradiance to match that of an epithelial cell such that the acquisition times and S/N ratio values would relate to biomedical applications. Following this, the algorithm was applied to spectra recorded from three different cell groups: mesenchymal stem cells and their vascular and osteogenic progeny. For further details on cultivation and preparation of these cells please refer to Molony et al.[35]

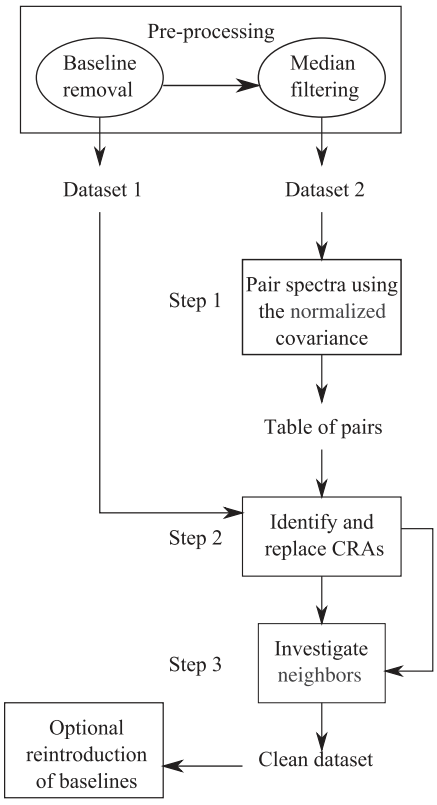

Figure 1. A flow chart of the proposed algorithm with additional pre-processing steps to deal with varying baselines.

Recording Spectra

A custom-built confocal Raman micro-spectrometer was used to record spectra from the polymer material. This system uses a 150 mW 532 nm laser and a diffraction grating with 600 lines/mm. More details on the specific system can be found in Kerr et al.[32] A sufficiently defocused, low numerical aperture microscope objective (Olympus UMplanFl 4x/0.1) was used in order to produce spectra from the polymer material that had a S/N ratio equivalent to that expected from an epithelial cell using a commercial Raman micro-spectrometer over a 60 s acquisition time. Maximum cooling of the camera (Andor Newton 920 BVF) was used in order to minimize dark current noise. The low magnification of the microscope objective provides for a large depth of field, which prevents any major change in focus over the course of the experiment, further reducing experimental variability across the acquired data sets. A single acquisition data set of 100 spectra was acquired with a 60 s integration time and a double acquisition data set of 2×100 spectra was acquired, each with a 30 s integration time so that a comparison of the proposed algorithm to the double acquisition method could be made in the context of S/N ratio. For the cell spectra, a commercial Raman microspectrometer was employed also using a 532 nm laser source. More information on this system is found in Molony et al.[35]

Measuring Signal-to-Noise Ratio

Experimentally, the S/N ratio can be estimated by the ratio of the intensity of the highest peak to the standard deviation of the noise in the spectrum.[36] The noise signal is isolated by performing a least squares fit of a reference spectrum to the denoised data set and subtracting this reference from the fitted spectrum. The reference may be the mean spectrum of a suitably large data set or may be obtained from a relatively low noise spectrum acquired over a long acquisition time that provides an accurate representation of the true irradiance. For the results presented in the Results section, the reference is taken to be the mean spectrum for a given data set. The intensity of the highest peak is taken from the fitted reference spectrum rather than the raw spectrum, which may include a significant noise component at that point, and would, therefore, affect the measurement of the S/N ratio. This process is illustrated in Fig. 2.

Results

Application to Polymer Data

The spectral data sets from the polymer material were processed using both the proposed algorithm and the double exposure method. The resulting CRA removed spectra were examined and compared in terms of S/N ratio using the approach described in the Measuring Signal-to-Noise Ratio section. Figure 3 illustrates the raw data, the removed cosmic rays, and the denoised data set following processing with both methods. Both algorithms make negligible changes to the underlying spectrum, aside from the areas contaminated with CRAs while retaining a high sensitivity for low intensity and broad CRAs. Figure 4 shows a magnified region (825–975 cm^{-1}) of the spectra to further illustrate the effectiveness of the method. This region was chosen in order to illustrate the algorithms ability to discriminate between spectral features and CRAs as it contains a number of peaks that vary in width and height. While both methods perform similarly in terms of CRA removal, there is, however, a difference in the S/N ratio of the denoised spectra obtained using the two methods. It should also be noted that in the denoised data set of the double acquisition method illustrated in Fig. 3, there are

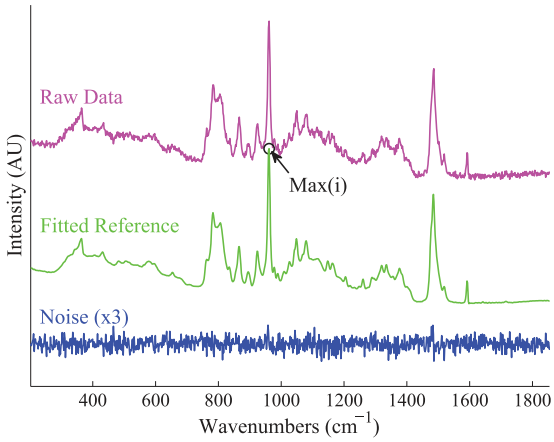

Figure 2. An illustration of the S/N ratio calculation.

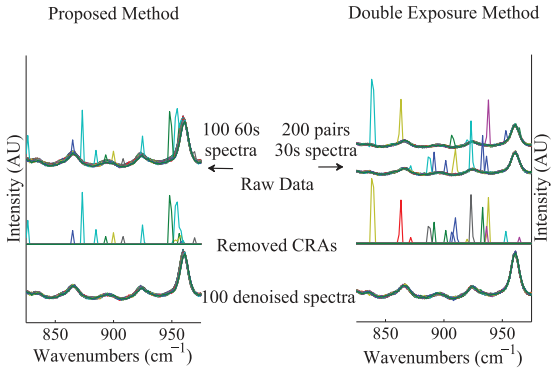

Figure 4. A magnified region of Fig. 3 comparing the raw data, clean data set, and the difference spectra to further illustrate the operation of the proposed algorithm.

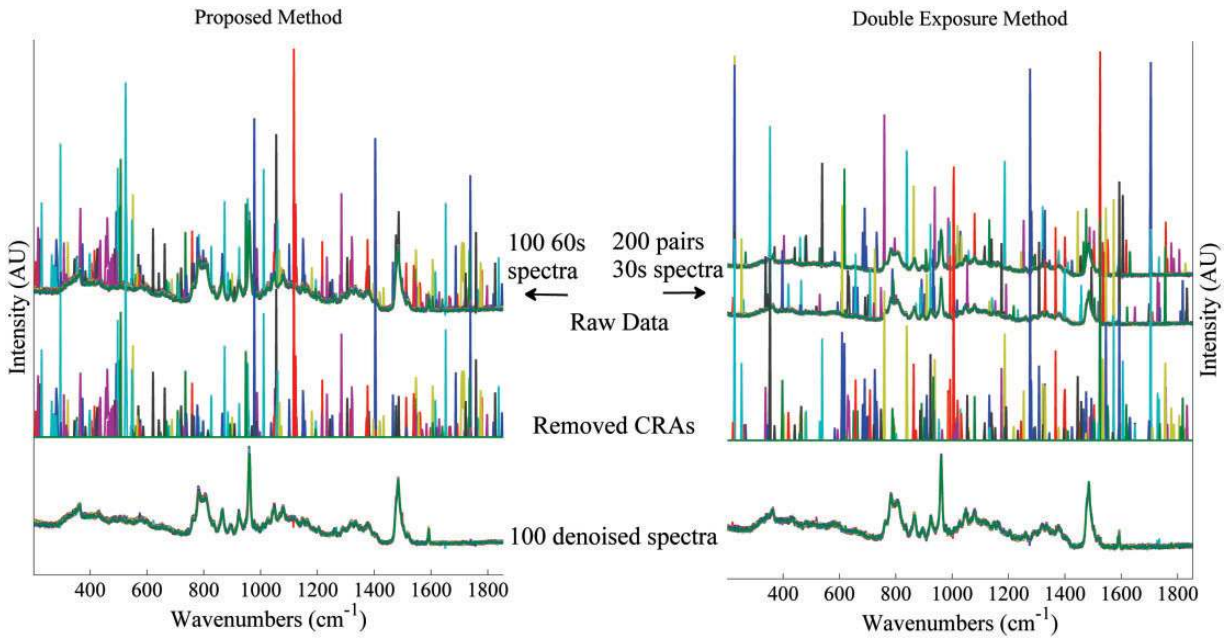

Figure 3. Illustration of the data sets used to evaluate the performance of the proposed algorithm in terms of S/N ratio. Left: A single data set of 100 raw spectra is shown with an acquisition time of 60 s. Right: Two data sets of consecutively collected spectral pairs with an acquisition time of 30 s. In both cases the raw data, removed CRAs, and denoised data set are shown.

the remnants of two CRAs evident near 1750 cm^{-1}. These are the remnants of two intense cosmic rays that were spread over multiple pixels. The outer edges of these CRAs were small enough to fall under the designated threshold. In cases such as this, investigating the neighboring pixels of identified CRAs with a lower threshold is necessary.

The mean spectrum of all 300 spectra collected in the experiment was used as the reference spectrum for measuring the S/N ratio as described in the previous section. Figure 5 illustrates the S/N ratio calculated over the data set of 100 denoised spectra for both the proposed algorithm and the double acquisition method. Of the data set that is denoised by the proposed algorithm, the range of S/N ratio values is 98–117 with the central 50% of S/N ratio values in the range of 104–111. For the data set that is denoised by the double acquisition method, the range of S/N ratio values is 90–104 with the central 50% of S/N ratio values in the range of 96–99. More than 75% of the denoised spectra processed using the proposed method exhibit higher S/N ratio values than those denoised using the double acquisition method.

Application to Biological Data

In order to evaluate the proposed algorithm's performance on biological spectra, three data sets recorded from three different cell groups were amalgamated into a single data set to which the algorithm was applied. Three data sets were recorded from: (1) mesenchymal stem cells (MSC), (2) the vascular progeny of MSC samples, and (3) the osteogenic progeny of MSC samples. It should be noted that the osteogenic cells contain a noticeable difference to the other samples, specifically the peak at 960 cm^{-1} that indicates the presence of phosphates. Both pre-processing steps were applied as illustrated in Fig. 1. A mean spectrum taken from the entire data set was used in the background subtraction algorithm and a polynomial of order 5 was also used to remove varying baselines in the data set. This fitted data set was then filtered using a median filter of size 11 before applying Step 1 of the proposed algorithm. Figure 6 illustrates the data sets at different stages of the CRA removal algorithm from raw spectra in Fig. 6a to the final data set without CRAs in Fig. 6d. Figure 6b shows the raw data following background subtraction. The resulting data set is then CRA removed using the proposed algorithm to produce the denoised data set shown in Fig. 6c. Finally, the background components are reintroduced to each individual spectrum.

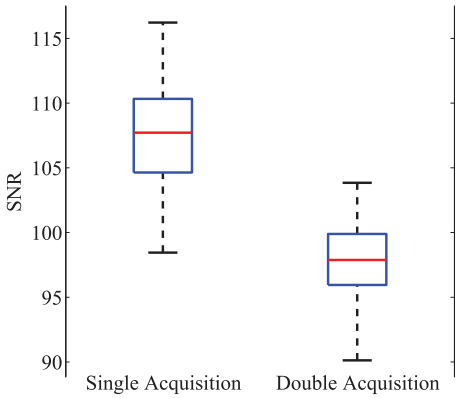

Figure 5. A boxplot of the resulting S/N ratios of the CRA removed spectra of both the proposed algorithm and the double exposure method.

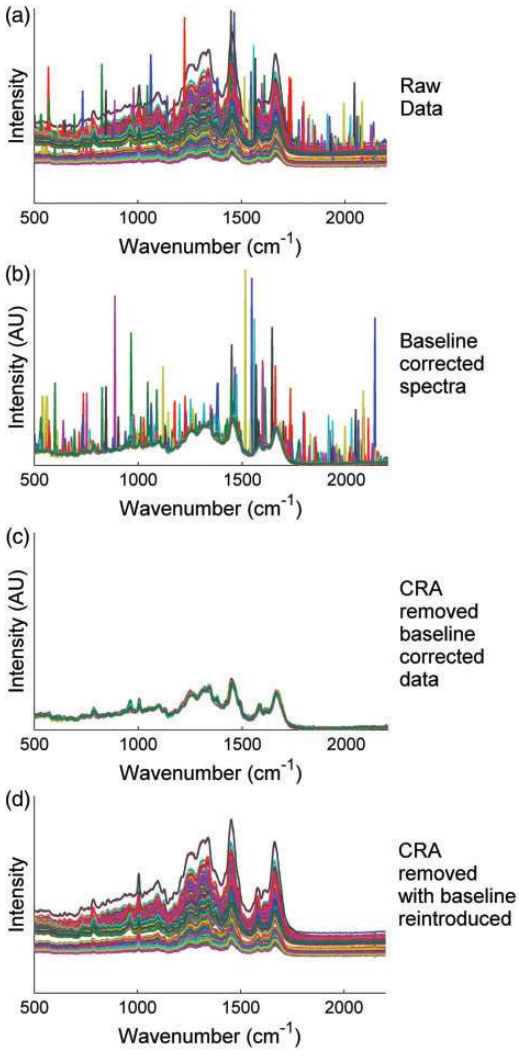

Figure 6. An illustration of the CRA removal of the data set. The raw data are background subtracted, CRA removed, and, finally, the background is then reintroduced to the data. The *y*-axis is fixed to the same values, for all figures.

Conclusion

Cosmic ray artifacts can be removed from spectra using a number of different methods. These include algorithms that can be directly applied to a single recorded spectrum using some form of digital filtering; however, such algorithms have the advantage of being applicable to dynamically changing samples, the goal of removing the cosmic ray while making no other change to the spectrum is challenging. A second group of algorithms for the removal of cosmic rays involves the capture of successive spectra from a sample that is not expected to change between captures. A direct comparison of subsequent spectra allows for the accurate removal of cosmic rays while making little or no other alteration to the underlying spectrum. The proposed algorithm relates to both of these approaches; the algorithm requires only a single recorded spectrum so long as a data set of similar spectra is available, a requirement that is naturally met for a large number of applications that involve the repeated capture of data.

In this paper, it has been demonstrated that the proposed algorithm does not require any a priori knowledge of system and camera parameters that were used to record the spectrum. In terms of effectiveness at CRA removal, this method performs similarly to the double acquisition method,[17] which is widely applied in the field of Raman spectroscopy. However, the algorithm differs from traditional sequential scan methods in that it incorporates a comparison across a data set of similar spectra, which is similar to NNC for point scanning spectroscopy. This results in an algorithm that combines aspects of both categories one and two, as described in the Background section.

For those cases where the amplitude of the shot noise and camera dark current dominates, this difference in S/N ratio between a single and a double acquisition may be negligible; however, in applications where low intensity spectra are collected or high read-out rate are required, this additional noise may become a significant factor and negatively affect the S/N ratio. The proposed algorithm has the advantage that it does not require the repeated capture of spectra and has shown that an overall improvement of S/N ratio of 10% can be expected for the recording conditions associated with biological samples. If desired, this may be translated into a decrease in acquisition time due to the square root relationship between intensity and the noise. Disregarding read noise and dark current a 10% improvement in S/N ratio translates to approximately a 20% reduction in acquisition time to obtain the same S/N ratio.

A second advantage of the proposed algorithm over the double acquisition method is that databases of previously recorded spectra can also be processed. It is notable that the algorithm is able to successfully pair the spectra within the data set despite the presence of spectra from three distinct cell groups. It can be expected that this feature may be extended to data sets containing a large number of spectra originating from disparate sources.

It must be acknowledged that the proposed algorithm will fail if a recorded spectrum contains legitimate spectral peaks that are unique to the data set that is employed for cosmic ray removal; in such a case, such peaks would be deemed to be cosmic rays and removed. However, for many applications of spectroscopy, and for a sufficiently large data set, the probability of such an occurrence can be expected to be low.

Acknowledgments

The authors thank the IRC and SFI for their support as well as their collaborator, Claire Molony, for recording the data sets of biological Raman spectra.

Conflict of Interest

The authors report there are no conflicts of interest.

Funding

This research was conducted with the financial support of the Irish Research Council (IRC) under project ID GOIPG/2013/1434 and Science Foundation Ireland (SFI) under Grant Number 15/CDA/3667.

ORCID iD

Sinead J. Barton http://orcid.org/0000-0003-4915-7335

References

1. M. Pelletier. "Quantitative Analysis Using Raman Spectrometry". Appl. Spectrosc. 2003. 57(1): 20A–20A.
2. D. Groom. "Cosmic Rays and Other Nonsense in Astronomical CCD Imagers". Exp. Astro. 2002. 14(1): 45–55.
3. L. Zhang, M. Henson. "A Practical Algorithm to Remove Cosmic Spikes in Raman Imaging Data for Pharmaceutical Applications". Appl. Spectrosc. 2007. 61(9): 1015–1020.
4. N. Stone, C. Kendall, J. Smith, et al. "Raman Spectroscopy for Identification of Epithelial Cancers". Farad. Discuss. 2004. 126: 141–157.
5. S. Dochow, N. Bergner, C. Krafft, et al. "Classification of Raman Spectra of Single Cells with Autofluorescence Suppression by Wavelength Modulation". Anal. Meth. 2013. 5(18): 4608–4614.
6. S. Li, L. Dai. "An Improved Algorithm to Remove Cosmic Spikes in Raman Spectra for Online Monitoring". Appl. Spectrosc. 2011. 65(11): 1300–1306.
7. F. Ehrentreich, L. Sümmchen. "Spike Removal and Denoising of Raman Spectra by Wavelet Transform Methods". Anal. Chem. 2001. 73(17): 4364–4373.
8. M. Soneira, R. Perez-Pueyo, S. Ruiz-Moreno. "Raman Spectra Enhancement with Fuzzy Logic Approach". J. Ram. Spectrosc. 2002. 33(8): 599–603.
9. Y. Katsumoto, Y. Ozaki. "Practical Algorithm for Reducing Convex Spike Noises on a Spectrum". Appl. Spectrosc. 2003. 57(3): 317–322.
10. W. Hill, D. Rogalla. "Spike-Correction of Weak Signals from Charge-Coupled Devices and its Application to Raman Spectroscopy". Anal. Chem. 1992. 64(21): 2575–2579.
11. G. Phillips, J. Harris. "Polynomial Filters for Data Sets with Outlying or Missing Observations: Application to Charge Coupled Device Detected Raman Spectra Contaminated by Cosmic Rays". Anal. Chem. 1990. 62(21): 2351–2357.

12. L. Quintero, S. Hunt, M. Diem. "Denoising of Raman Spectroscopy Signals". Poster presented at the 2006 Thrust R2C Multi Spectral Discrimination Methods Conference, Bernard M. Gordon Center for Subsurface Sensing and Imaging Systems (Gordon-CenSSIS), Boston, MA; January 1, 2006. http://hdl.handle.net/2047/d10008330 [accssed Apr 2 2019].
13. A. Maury, R.I. Revilla. "Autocorrelation Analysis Combined with a Wavelet Transform Method to Detect and Remove Cosmic Rays in a Single Raman Spectrum". Appl. Spectrosc. 2015. 69(8): 984–992.
14. D. Zhang, K. Jallad, D. Ben-Amotz. "Stripping of Cosmic Spike Spectral Artifacts Using a New Upper-Bound Spectrum Algorithm". Appl. Spectrosc. 2001. 55(11): 1523–1531.
15. D. Zhang, D. Ben-Amotz. "Removal of Cosmic Spikes from Hyper-Spectral Images Using a Hybrid Upper-Bound Spectrum Method". Appl. Spectrosc. 2002. 56(1): 91–98.
16. D. Zhang, J. Hanna, D. Ben-Amotz. "Single Scan Cosmic Spike Removal Using the Upper Bound Spectrum Method". Appl. Spectrosc. 2003. 57(10): 1303–1305.
17. H. Takeuchi, I. Harada. "Simple and Efficient Method to Eliminate Spike Noise from Spectra Recorded on Charge-Coupled Device Detectors". Appl. Spectrosc. 1993. 47(1): 129–131.
18. S.M. Anthony, J.A. Timlin. "Removing Cosmic Spikes Using a Hyperspectral Upper-Bound Spectrum Method". Appl. Spectrosc. 2017. 71(3): 507–519.
19. S. Mohzharov, A. Nordon, D. Littlejohn, et al. "Automated Cosmic Spike Filter Optimized for Process Raman Spectroscopy". Appl. Spectrosc. 2012. 66(11): 1326–1333.
20. W. Chew. "Information-Theoretic Chemometric Analyses of Raman Data for Chemical Reaction Studies". J. Ram. Spectrosc. 2011. 42(1): 36–47.
21. T.M. James, M. Schlösser, R.J. Lewis, et al. "Automated Quantitative Spectroscopic Analysis Combining Background Subtraction, Cosmic Ray Removal, and Peak Fitting". Appl. Spectrosc. 2013. 67(8): 949–959.
22. J. Zhao. "Image Curvature Correction and Cosmic Spike Removal Removal for High-Throughput Dispersive Raman Spectroscopy". Appl. Spectrosc. 2003. 57(11): 1368–1375.
23. U. Cappel, I. Bell, L. Pickard. "Removing Cosmic Ray Features from Raman Map Data by a Refined Nearest Neighbor Comparison Method as a Precursor for Chemometric Analysis". Appl. Spectrosc. 2010. 64(2): 195–200.
24. C. Behrend, C. Tarnowski, M. Orris. "Identification of Outliers in Hyperspectral Raman Image Data by Nearest Neighbour Comparison". Appl. Spectrosc. 2002. 56(11): 1458–1461.
25. B. Li, A. Calvet, Y. Casamayou-Boucau, et al. "Kernel Principal Component Analysis Residual Diagnosis (Kpcard): An Automated Method for Cosmic Ray Artifact Removal in Raman Spectra". Anal. Chim. Act. 2016. 913: 111–120.
26. P. Crow, A. Molckovsky, N. Stone, et al. "Assessment of Fiberoptic Near-Infrared Raman Spectroscopy for Diagnosis of Bladder and Prostate Cancer". Urology. 2005. 65(6): 1126–11230.
27. P.R. Jess, D.D. Smith, M. Mazilu, et al. "Early Detection of Cervical Neoplasia by Raman Spectroscopy". Int. J. Canc. 2007. 121(12): 2723–2728.
28. A. Haka, K. Shafer-Peltier, M. Fitzmaurice, et al. "Identifying Microcalcifications in Benign and Malignant Breast Lesions by Probing Differences in their Chemical Composition Using Raman Spectroscopy". J. Canc. Res. 2002. 62(18): 5375–5380.
29. D. Dussault, P. Hoess P. "Noise Performance Comparison of ICCD with CCD and EMCCD Cameras". Proc. SPIE Int. Soc. Opt. Eng. 2004. 5563: 195–204.
30. S. Barton, B. Hennelly. "Signal to Noise Ratio of Raman Spectra of Biological Samples". Proc. SPIE Int. Soc. Opt. Eng. 2018. 10685: 106854F.
31. N. Afseth, A. Kohler. "Extended Multiplicative Signal Correction in Vibrational Spectroscopy, A Tutorial". Chemom. Int. Lab. Sys. 2012. 117: 92–99.
32. L. Kerr, H. Byrne, B. Hennelly. "Optimal Choice of Sample Substrate and Laser Wavelength for Raman Spectroscopic Analysis of Biological Specimen". Anal. Meth. 2015. 7(12): 5041–5052.
33. B.D. Beier, A.J. Berger. "Method for Automated Background Subtraction from Raman Spectra Containing Known Contaminants". Analyst. 2009. 134(6): 1198–1202.
34. D. Liu, H. Byrne, L. O'Neill, et al. "Investigation of Wavenumber Calibration for Raman Spectroscopy Using a Polymer Standard". Proc. SPIE Int. Soc. Opt. Eng. 2018. 10680: 1080627.
35. C. Molony, J. McIntyre, A. Maguire, et al. "Label-Free Discrimination Analysis of De-differentiated Vascular Smooth Muscle Cells, Mesenchymal Stem Cells and their Vascular and Osteogenic Progeny Using Vibrational Spectroscopy". Biochim. Biophys. Acta Mol. Cell. Res. 2018. 1865(2): 343–353.
36. T.J. Harvey, C. Hughes, A.D. Ward, et al. "Classification of Fixed Urological Cells Using Raman Tweezers". J. Biophot. 2009. 2(1–2): 47–69.