

## ALGORITHMS OF CHEMICALS DETECTION USING RAMAN SPECTRA

**Andrzej Kwiatkowski, Marcin Gnyba, Janusz Smulko, Paweł Wierzba**

Gdansk University of Technology, Faculty of Electronics, Telecommunications and Informatics, G. Narutowicza 11/12, 80-233 Gdansk, Poland (akwiatkowski@zak.eti.pg.gda.pl, mgnyba@eti.pg.gda.pl, ✉ jsmulko@eti.pg.gda.pl +48 58 348 6095, pwierzba@eti.pg.gda.pl)

### Abstract

Raman spectrometers are devices which enable fast and non-contact identification of examined chemicals. These devices utilize the Raman phenomenon to identify unknown and often illicit chemicals (*e.g.* drugs, explosives) without the necessity of their preparation. Now, Raman devices can be portable and therefore can be more widely used to improve security at public places. Unfortunately, Raman spectra measurements is a challenge due to noise and interferences present outside the laboratories. The design of a portable Raman spectrometer developed at the Faculty of Electronics, Telecommunications and Informatics, Gdansk University of Technology is presented. The paper outlines sources of interferences present in Raman spectra measurements and signal processing techniques required to reduce their influence (*e.g.* background removal, spectra smoothing). Finally, the selected algorithms for automated chemicals classification are presented. The algorithms compare the measured Raman spectra with a reference spectra library to identify the sample. Detection efficiency of these algorithms is discussed and directions of further research are outlined.

Keywords: data processing algorithms, noise, interferences, Raman spectroscopy.

© 2010 Polish Academy of Sciences. All rights reserved

### 1. Introduction

The Raman phenomenon was discovered in 1928 by C.V. Raman. He observed the appearance of photons having wavelengths different than that of the incident beam, when the investigated material was illuminated by an intense beam of monochromatic light (Fig. 1). These photons result from non-elastic scattering of the photons from the incident beam on the molecules of the substance. The spectrum of the reflected photons contains distinctive bands, whose wavelengths and relative intensities are unique to individual substances or mixtures, thus allowing their identification.

At present, Raman spectrometers of different size and accuracy are built. During the global threat of terrorism, such portable devices may be used at airports, railway stations or other public buildings to improve the security level. These devices can easily and almost instantly identify the suspected substances even by semiskilled staff. Unfortunately, a portable Raman spectrometer has numerous drawbacks when compared with its laboratory version.

The spectrometer utilizes a laser with a narrow spectral line and sufficiently low inherent noise component to irradiate effectively the inspected samples. The semiconductor lasers of 785 nm wavelength with optical output power of a few hundred mW are commonly used in portable devices. The scattered light passes through an optical system that commonly comprises an optical filter to minimize back-reflected laser light and Rayleigh scattering signal as well as a diffraction grating which disperses scattered light into single wavelength components. Dispersed components are directed onto a CCD matrix having sufficient resolution (number of pixels), that collects a charge proportional to the intensity of the scattered light. The detected light is significantly less intense than the laser light and therefore the measurement results are very sensitive to various interferences and inherent system noise.

The excitation laser and CCD matrix are connected through a control unit to a computer with a dedicated software which allows signal acquisition and preprocessing as well as proper detection of the chemicals. Identification of the irradiated sample is performed by comparing the measured spectrum with the spectra stored in the database.

Such device is currently developed at the Faculty of Electronics, Telecommunications and Informatics, Department of Optoelectronics and Electronics Systems, Gdańsk University of Technology (Fig. 2). The device can work with two lasers having different wavelengths: 785 nm (made by Ocean Optics) and 355 nm (made by Cobolt AB). The latter laser has lower optical output power than the former one, but it is offset by stronger Raman effect at UV excitation. The dedicated spectrometers were produced by Ocean Optics. The distance between the probe and the sample depends on the probe focus and is usually around a few cm. Resolution of the registered Raman spectra depends on Raman shift and is close to  $10 \text{ cm}^{-1}$ . Spatial resolution of the measurements is limited by the laser spot and equals approx 1–2 mm.

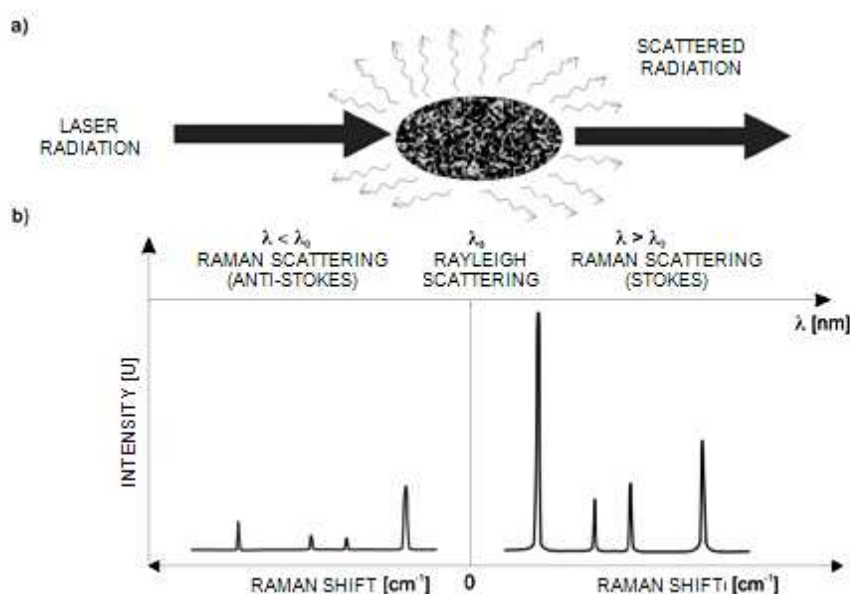


Fig. 1. Illustration of the Raman phenomenon: a) scattering of laser radiation; b) Raman spectrum.

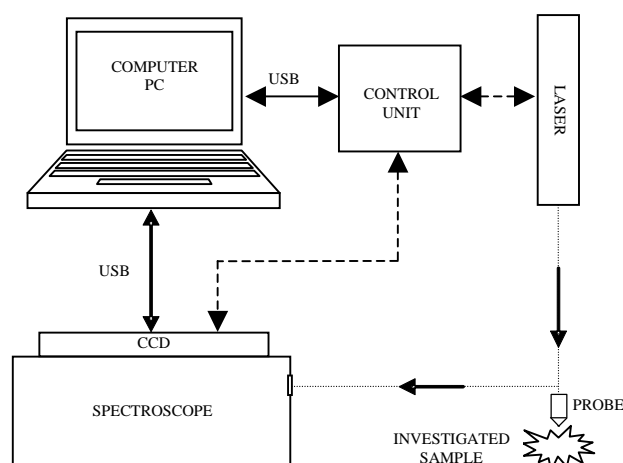


Fig. 2. Block diagram of a portable Raman spectrometer.

The dedicated control unit of the measurement system is attached to a PC computer through an USB interface. This unit controls the lasers' output power and triggers the

spectrometer.

The laser can work in constant mode of optical output power (CW) or with the output power modulated by harmonic or rectangular signals for synchronous detection that reduces the influence of external interferences. In the latter mode, an additional algorithm, used in the lock-in amplifiers, has to be applied to recover the amplitude of the detected signal. The control unit manages also CCD matrix cooling, which is used to reduce the thermal noise of the photodetector. Additionally, the control unit switches the lasers and signalizes their work due to security standards for class IIIb lasers [1].

## 2. Noise and interferences in measurements

The main sources of interference during Raman spectra measurements in the field are: fluorescence induced by the laser beam, cosmic radiation, external light sources and noise of the CCD matrix and applied charge amplifiers [2]. Influence of these factors can be reduced only partially by various methods using more expensive and sophisticated equipment that consumes additional energy.

Fluorescence is the emission of light by the irradiated substance in a range of wavelengths longer than those of the excitation source. This effect is due to complex electron transitions between energy levels, associated with light absorption, non-emission transitions and light emission when the electrons return to lower energy levels.

Another source of interference present in Raman spectra is cosmic radiation, which can form narrow spikes (much more narrow than the peaks caused by Raman phenomena). Cosmic radiation exhibits a noise component present in the whole range of Raman spectra.

The inherent noise of a Raman spectrometer is determined by thermal noise of the CCD matrix detector. That detector records the scattered Raman radiation whose intensity can be very low – tens of photons per second. In such circumstances thermal noise of the CCD detector and of the charge amplifiers can obscure the right signal. To reduce this effect, the CCD matrix has to be cooled. The portable Raman spectrometer due to limited dimensions and energy supplies uses thermoelectric cooling (*i.e.* a Peltier module).

## 3. Methods of Raman spectra preprocessing

Interference and errors of the recorded spectra limit proper identification of the examined samples. Therefore, each spectrum should be pre-processed to remove the disturbing components (*e.g.* spikes, fluorescence background, noise component) before applying detection algorithms. The portable Raman spectrometer needs a dedicated software that can perform measurements and further processing at a very limited operator involvement and computing complexity.

The process of Raman spectra recognition comprises three main stages (Fig. 3). Initially, the background has to be removed from the spectrum. This component is the result of fluorescence, background noise of the CCD detector and other light sources present during field measurements [1]. The background of the spectrum is approximated by a given mathematical function and then the background is subtracted from the measured spectrum. The simplest method is the use of a first-degree polynomial:

$$y = Mx + B, \quad (1)$$

where:

$$M = \frac{y_2 - y_1}{x_2 - x_1}, \quad B = y_1 - M \cdot x_1. \quad (2)$$

Values  $x_1$ ,  $x_2$ ,  $y_1$ ,  $y_2$  mean the coordinates of two points which belong to the measured spectrum and are arbitrarily chosen by the operator (e.g. two points being placed close to the extreme values of the measured Raman shift – Fig. 4).

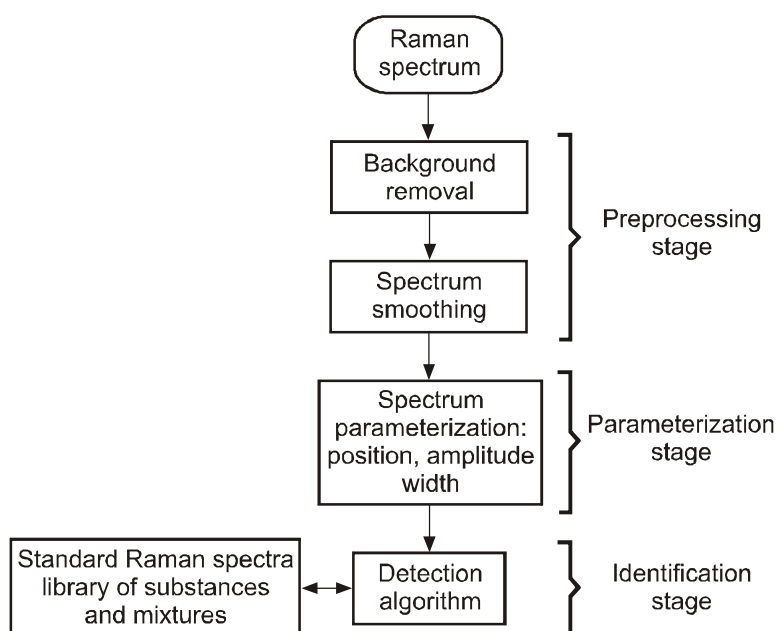


Fig. 3. Block diagram of Raman spectra processing during chemicals detection.

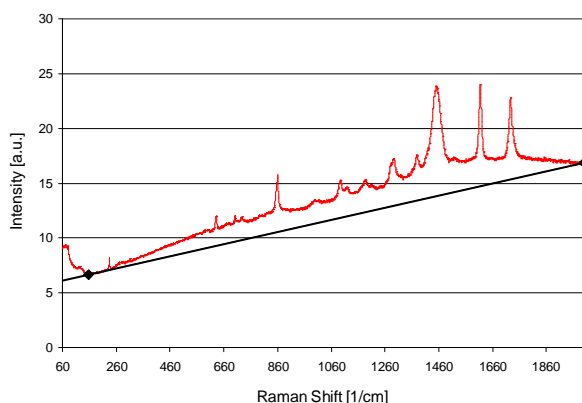


Fig. 4. Approximation of the background using a line for two arbitrarily chosen points.

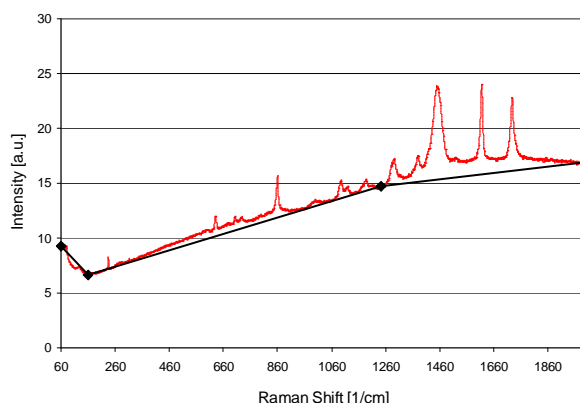


Fig. 5. Approximation of the background using four points.

That algorithm can be extended simply by a baseline approximation which uses a broken line at the points indicated by the device operator. This method gives usually better results than that previously presented (Fig. 5), but cannot be done automatically without operator attention. Good results are observed even if a second-order polynomial is used (Fig. 6). Polynomials of a degree higher than three are not recommended due to possible adverse waving.

A method of background removal without any operator's intervention is the GIFTS algorithm [3–6]. The baseline is determined by using repeatedly the least-squares method. In the first step the straight line is established by the entire measured spectrum and this is a comparison threshold for the subsequent stages. The next step is to count the points of the spectrum located above and below the threshold line. If there are fewer points above the line than below, they are considered peaks and are replaced with an approximating straight line. In

the next iteration a new line of the threshold is calculated and points are counted and compared again. The computing process is repeated until the number of points above the new threshold is less than or equal to the number of points below this threshold. The straight line obtained in the last iteration is considered as a baseline and subtracted from the originally measured spectrum. Fig. 7 illustrates the results of use of this algorithm.

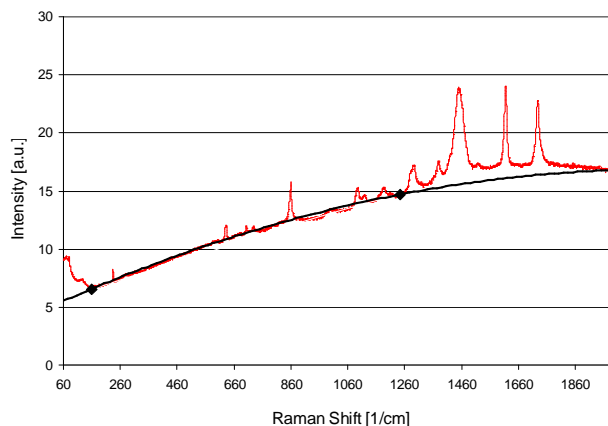


Fig. 6. Background approximation using a second order polynomial.

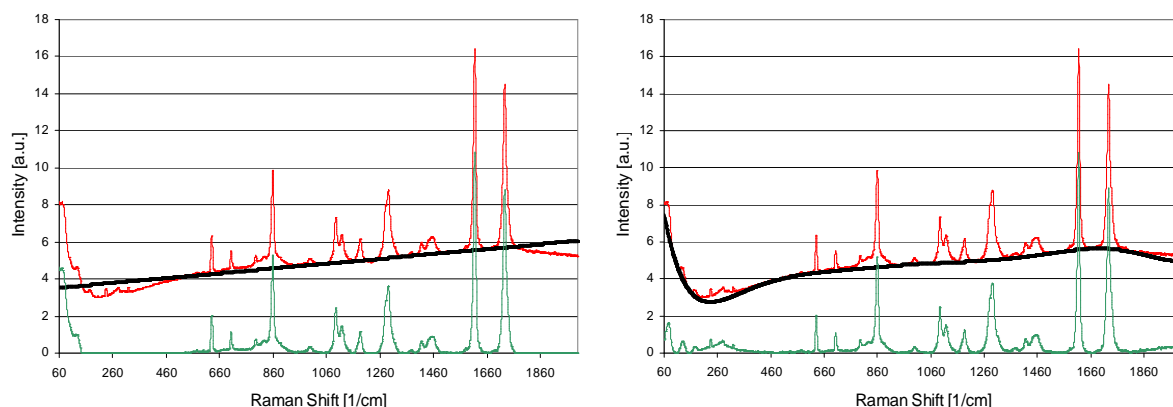


Fig. 7. Result of the GIFTS algorithm (left) and an iterative polynomial fitting (right): approximated background (bold line) and spectrum after background removal (gray line).

In some cases it may happen that the algorithm cannot converge on an exact result. Then the algorithm is terminated after the assumed number of iterations. The presented method is effective to remove the background to the Raman spectra with positive peaks.

The automatic baseline correction can be done with iterative polynomial fitting (Fig. 7). The points placed above the polynomial are removed in each iteration. In this way, a new spectrum is created which will be used in the next iteration. The algorithm stops after a fixed number of iterations (*e.g.* 30).

The second step of pre-processing – spectrum smoothing – is performed to reduce its random component and accidental spikes. There are various algorithms proposed in literature (*e.g.* Fourier or wavelet transform denoising), but a Savitzky and Golay smoothing filter is a good choice [7].

The Savitzky and Golay method smoothes the noise component using piece-by-piece fitting of a polynomial function to the right signal. The fitting is done by minimization of the least squares measure. The width of the window and the degree of the polynomial is fixed. Fig. 8 shows the results of the smoothing algorithm using a 10-point wide window and a second-order polynomial for approximation. The achieved results prove that the method can

effectively reduce random components of the Raman spectrum, while preserving the original shape of the spectral lines.

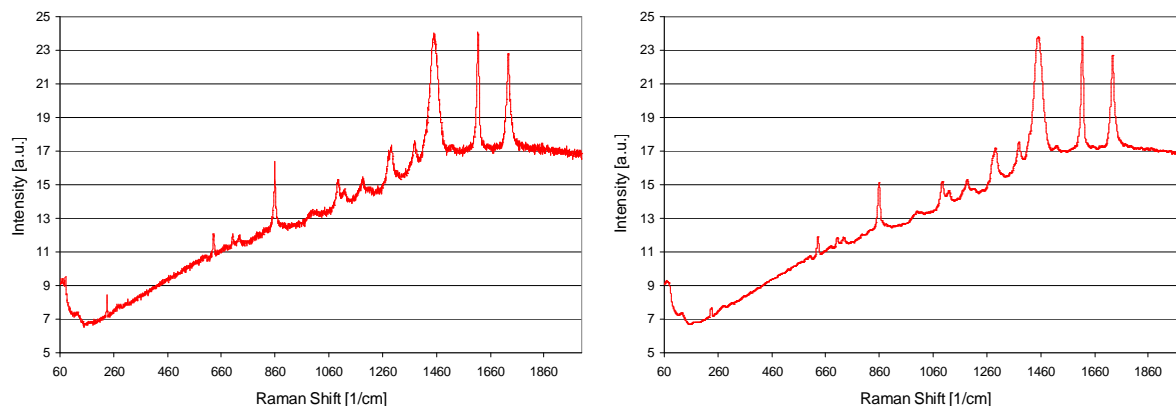


Fig. 8. The smoothed Raman spectrum (left) obtained from the measured one by applying the Savitzky-Golay smoothing filter (right).

The pre-processed spectrum is parameterized to establish a set of parameters (positions of the spectra peaks, their relative amplitudes and their widths) that give essential information necessary for detection of the investigated chemicals. Identification of the spectral lines is the main objective of Raman spectra analysis because this leads directly to identification of the examined chemicals. The pre-processed Raman spectra can identify the chemicals by using one of the several and well-known methods [6]. The correlation methods, rating similarity of the registered and pre-processed spectrum with the library reference spectra, are applied. Other methods identify each spectrum line characteristic for the investigated substance [8–10]. Thereby, instead of the whole Raman spectrum, a reduced set of numbers has to be preserved and analyzed. Usually, only three parameters (peak value, the width and the position) of every spectral line are analyzed. A set of the reduced data (Fig. 9) is the input for the detection algorithms.

The spectral line shapes are described by a Lorentz function, Gaussian function or with the mixed Gaussian- Lorentz function according to the following formulas:

$$f_L(v) = \frac{A \cdot \Delta v^2}{(v - v_0)^2 + \Delta v^2}, \quad (4)$$

$$f_G(v) = A \cdot \exp\left[-\frac{(v - v_0)^2}{2 \cdot \Delta v^2}\right], \quad (5)$$

$$f(v) = (100\% - M) \cdot f_G(v) + M \cdot f_L(v), \quad (6)$$

where:  $f_L(v)$  – Lorentz function;  $f_G(v)$  – Gaussian function;  $f(v)$  – mixed Gaussian-Lorentz function;  $M$  – mixing ratio [%];  $A$  – peak's value of the spectral line;  $\Delta v$  – half of FWHM (Full Width at Half Maximum);  $v_0$  – position of the spectral line.

To find positions of the spectral lines a derivative of the spectrum is examined. This function changes the sign from positive to negative at a point of the local maximum. This point corresponds to a frequency of the spectral line and is marked as  $v_0$ . When the peak of the spectral line lies between the points at which the Raman spectrum was measured, the exact value of  $v_0$  is computed by using a second order polynomial approximation of  $N$  points placed

symmetrically around its vicinity. For the appointed coefficients  $a_i$  the peak spectral line is counted from the formula:

$$A = \sum_{i=0}^2 a_i \cdot \nu_o^i \quad (7)$$

The FWHM is defined as the width of the spectral line at half of its peak value. The point coordinates which determine the FWHM value, can be computed by applying a straight line approximation between the points, being found above and below the half of the peak's value.

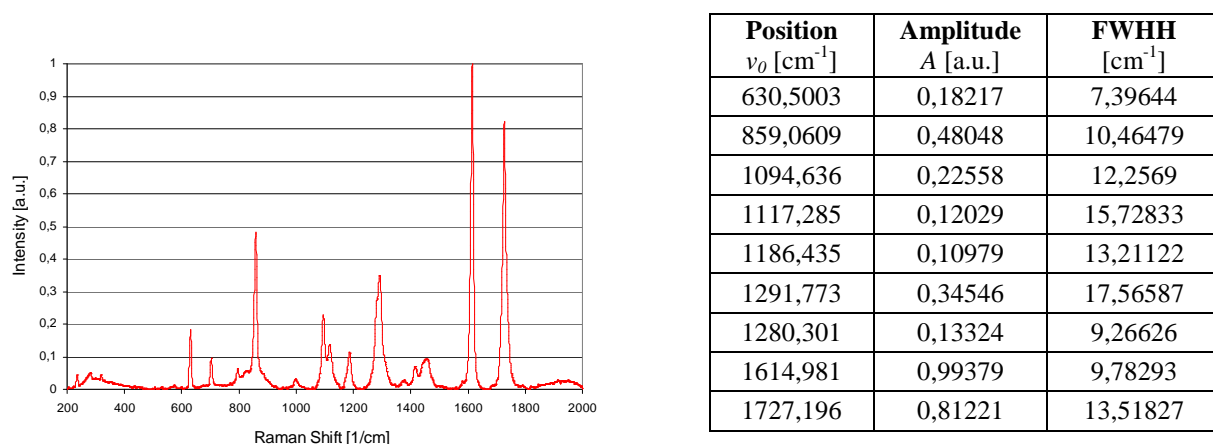


Fig. 9. The main parameters (right) of the exemplary Raman spectrum (left) estimated by the described algorithm.

#### 4. Effectiveness of detection algorithms

Chemicals are detected by comparing the registered spectrum with the reference spectra available in the database. The spectra are compared by summing up, according to different norms (e.g. absolute or squared values), the differences between the identified spectrum samples (or their derivatives) and the reference spectrum from the database. The outcome of such algorithms is a match quality  $M$  (mostly within the range 0÷100) which describes the degree of similarity between the compared spectra. The higher quantity means that the compared spectra are more similar. A few most popular algorithms are presented below. Each of the formulas marks  $s_i$  as the samples of the measured spectrum and  $r_i$  as the samples of the reference spectrum and computes the  $M$  value adequate for the given algorithm:

##### 1. Absolute Difference Value Search (ADV)

$$M_{ADV} = \left(1 - \frac{N}{D}\right) \cdot 100, \quad N = \sum_{i=1}^n |s_i - r_i|, \quad D = \sum_{i=1}^n |s_i|. \quad (8)$$

##### 2. First Derivative Absolute Value Search (FDAV)

$$M_{FDAV} = \left(1 - \frac{N}{D}\right) \cdot 100, \dots N = \sum_{i=1}^n |\Delta s_i - \Delta r_i|, \dots D = \sum_{i=1}^n |\Delta s_i|, \dots \begin{matrix} \Delta s_i = s_i - s_{i-1} \\ \Delta r_i = r_i - r_{i-1} \end{matrix} \quad (9)$$

##### 3. Least Squares Search (LS)

$$M_{LS} = \left(1 - \frac{N}{D}\right) \cdot 100, \quad N = \sum_{i=1}^n (s_i - r_i)^2, \quad D = \sum_{i=1}^n s_i^2 \quad (10)$$



4. First Derivative Least Square Search (**FDLS**)

$$M_{FDLS} = \left(1 - \frac{N}{D}\right) \cdot 100, \quad N = \sum_{i=1}^n (\Delta s_i - \Delta r_i)^2, \quad D = \sum_{i=1}^n \Delta s_i^2, \quad \begin{aligned} \Delta s_i &= s_i - s_{i-1} \\ \Delta r_i &= r_i - r_{i-1} \end{aligned} \quad (11)$$

5. Euclidean Distance Search (**ED**)

$$M_{ED} = \left(1 - \frac{N}{D}\right) \cdot 100, \quad N = \sum_{i=1}^n \sqrt{|s_i^2 - r_i^2|}, \quad D = \sum_{i=1}^n s_i. \quad (12)$$

6. Correlation Coefficient (**CC**)

$$M_{CC} = \frac{\sum_{i=1}^n (s_i - s)(r_i - r)}{\sqrt{\sum_{i=1}^n (s_i - s)^2 \sum_{i=1}^n (r_i - r)^2}} \cdot 100. \quad (13)$$

7. Correlation Search (**CO**)

$$M_{CO} = \frac{N}{D} \cdot 100, \quad N = \left(\sum_i \Delta s_i \Delta r_i\right)^2, \quad D = \sum_i \Delta s_i^2 \sum_i \Delta r_i^2, \quad \begin{aligned} \Delta s_i &= s_i - s_{i-1} \\ \Delta r_i &= r_i - r_{i-1} \end{aligned} \quad (14)$$

The presented algorithms are based on the measured spectra samples  $s_i$  without spectra parameterization, as presented earlier by peak positions or FWHM values.

To compare the efficiency of these algorithms, an exemplary Raman spectra database containing ten different substances was created. The spectra were measured in an acquisition time of 40 s and at different temperatures of the CCD matrix ( $-7^\circ\text{C}$  or  $-15^\circ\text{C}$ ). The reduced temperature of the CCD matrix below the ambient temperature limited thermal background noise and improved the efficiency of spectra identification. Each spectrum was submitted to the pre-processing stage. Then, the detection algorithms were applied to test their detection efficiency by comparing the database with another measured Raman spectrum for a mixture of 20% methanol and 80% ethanol using the same measurement conditions as for the database creation. The detailed results are shown in Table 1. The detected mixture (Chemical label 5, Table 1) was unambiguously identified by all considered algorithms but due to the measurement errors and unavoidable interferences the match quality  $M$  was lower than the maximum value of 100.

The same procedure was applied again but for the Raman spectrum of another substance – sugar (Chemical label 3, Fig. 10). Very similar conclusions can be deduced from the achieved results as those mentioned in the previous case presented in Table 1.

The general conclusion can be gathered that all algorithms performed the chemicals' detection correctly having the maximum value of match quality  $M$  for the detected substance. However, some of the algorithms had a reasonably high match quality for other substances that could lead to false detection when the samples are contaminated by ingredients. The correlation coefficient (CC) exhibited such feature. This algorithm detected correctly the proper chemical at a level closest to 100% when compared with the results obtained for other detection methods. Unfortunately, the CC algorithm demonstrated also a much higher degree of similarity with other chemical substances from the database than the remaining algorithms (Table 1). Such situation means excessive probability of wrong detection, especially in practice when the samples are usually mixed with some other chemicals of various characteristic or when measurements suffer from excessive interferences. Detection results of





such mixtures are very difficult to be predicted. We can certainly expect a decrease of detection efficiency that depends strongly on the type and amount of the additives.

Table1. Efficiency of detection algorithms obtained for a database containing ten substances.

Chemical name	Chemical label	CO	ADV	FDAV	LS	FDLS	CC	ED
Diamond	1	0	3,53	0	4,13	0	15,21	0
Isopropyl alcohol	2	1,94	30,69	0	44,17	0	55,62	8,56
Sugar	3	0	25,36	0	41,01	0	55,83	0
Tertensif SR	4	22,73	48,79	0	74,36	0	79,07	11,3
Methanol (20%) and ethanol (80%)	5	98,49	92,45	82,49	99,52	98,23	99,64	68,19
Acetylsalicylic acid	6	0,31	18,06	0	23,76	0	28,67	0
Ascorbic acid	7	5,83	0	0	13,85	0	70,93	0
Zyrtec UCB	8	0,64	15,68	0	18,32	0	20,25	0
Metronidazol	9	0,03	22,69	0	33,59	0	46,61	3,74
Polopiryna S	10	0,77	28,36	0	38,25	0	42,11	0,64

A similarly low selectivity of detection was observed for the algorithms LS and ADV (Table 1, Fig. 10). The results obtained by applying these algorithms were very sensitive to the efficiency of the spectrum background removal. It is observed that these algorithms are very vulnerable to the efficiency of the spectra background removal. The CO and ED algorithms are characterized by much better selectivity. The most selective (though not giving the maximum resemblance level) are algorithms where the first derivatives are compared (FDAV and FDLS). These algorithms are also more resistant to improper background removal. It is worth to mention that the CC and CO algorithms work correctly even without removal of the spectrum background.

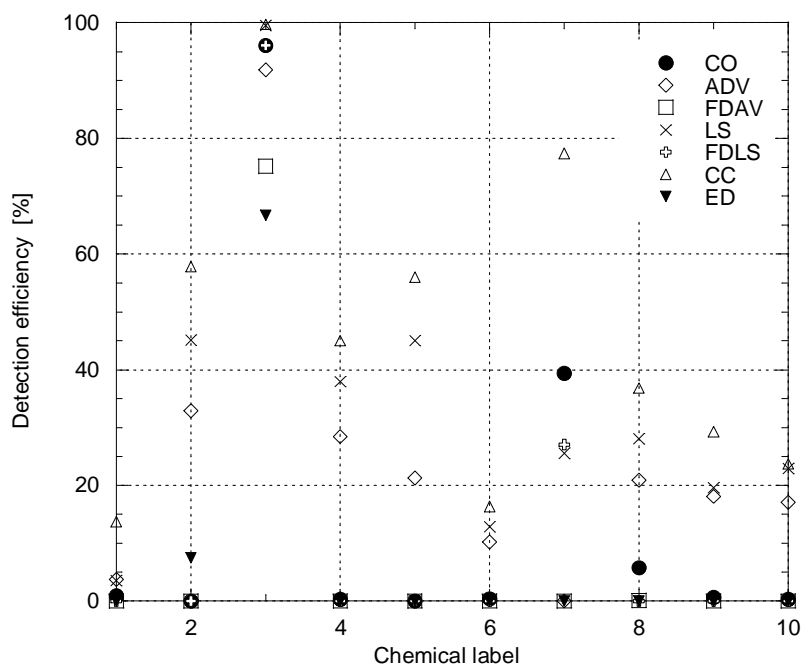


Fig. 10. Results of detection algorithms efficiency obtained for sugar identification (chemical label 3).

The data processing time of each algorithm is a very important parameter in portable devices. This is related directly to the requirements of energy consumption. Thus, to

characterize the selected algorithms, their time of computation (using an ordinary notebook computer working with Windows XP) was measured. The slowest algorithm was CO (Table 2). It is worth to underline that the algorithms (e.g. FDAV and FDLS) that have similar detection efficiency need completely different computation time. This effect is due to additional multiplication when the FDLS algorithm is applied. Thus, an optimal choice of the detection algorithm for the portable Raman spectrometer needs a thorough selection. We suppose that such selection should be preceded by careful analysis using a more abundant database. Such a database can be created by applying measurements from other devices after some necessary modifications [11]. Nevertheless, the results achieved for the database limited to ten spectra only can characterize the problem.

Measurements of Raman spectra can also identify various chemical substances composition. This is important in practice, especially at chemical reactions monitoring *in situ* in industry. Unfortunately, these possibilities are limited to some restricted cases when a mathematical model is established for Raman spectra measured for the mixtures with a fixed and known composition [8–10].

Table 2. Processing time of the applied detection algorithms for the measured Raman spectra using an ordinary notebook computer working with MS Windows XP operating system.

Algorithm name	Processing time [ms]
ADV	0,3991
ED	0,4423
FDAV	0,4757
LS	0,6461
CC	0,6951
FDLS	0,7177
CO	0,8507

## 5. Conclusions

This paper outlines details of the software dedicated for a portable Raman spectrometer. The stages of Raman spectra operations are presented with regard to measurements accuracy and effectiveness of the detection process. The introduced algorithms are selected to avoid user interventions during the measurements and identification process. We conclude that a portable Raman system can identify various chemicals very effectively by the chosen algorithms, well known from the literature. The algorithm has to be selected to maximize detection efficiency while having acceptable computation complexity. This selection is important to limit energy consumption of a portable device at similar detection efficiency. We conclude that the algorithms based on the first derivative of Raman spectra are most efficient for chemicals detection. In addition, the FDAV algorithm can be characterized by its limited processing time, that means low energy consumption.

## Acknowledgments

This research was financed by the Polish Ministry of Science and Higher Education (2009/2011), developmental project no. OR 0000 2907.

## References

- [1] European technical standard EN 60825-1:2007 Safety of laser products – Part 1: *Equipment classification and requirements*. Equivalen of the Polish technical standard PN-EN 60825-1: *Bezpieczeństwo urządzeń*



*laserowych – Część 1: Klasyfikacja sprzętu i wymagań.* (in Polish)

- [2] Bielecki, Z., Rogalski, A., (2001). *Detekcja sygnałów optycznych*. Warszawa: WNT. (in Polish)
- [3] Azarraga, L.V., Hanna D.A. *GIFTS, Athens ERL GC/FT-IR Software and User's Guide*.
- [4] Lam, T. (2004). A New Era in Affordable Raman Spectroscopy. *Raman Technology For Today's Spectroscopists*, 30-37.
- [5] McCreery R.L., *Raman Spectroscopy for Chemical Analysis*, John Wiley & Sons, New York 2000.
- [6] <https://ftirsearch.com/help/algo.htm>
- [7] Fochesatto, J., Sloan, J., (2009). Signal processing of multicomponent Raman spectra of particulate matter. Selected topics in electronics and systems. *World Scientific*, 49, 49-66,.
- [8] Næs, T., Isaksson, T., Fearn, T., Davies, T. (2002). *Multivariate Calibration and Classification*. NIR Publications.
- [9] Martens, H., Næs, T., (1989). *Multivariate Calibration*. John Wiley & Sons Ltd.
- [10] Cristianini, N., Shawe-Taylor, J. (2000). *An Introduction to Support Vector Machines*. Cambridge University Press.
- [11] Denton, M.B., Sperline, R.P., Giles, J.H., Gilmore, D.A., Pommier, C.J.S., Downs, R.T. (2003). Advances in the Application of Array Detectors for Improved Chemical Analysis, Part I. Comparison of Qualitative Analyses Using Large, Computer-Based Spectral Libraries. *Aust. J. Chem.*, 53, 117-131.

