Automated feature weighting and random pixel sampling in k-means clustering for Terahertz image segmentation

Mohamed Walid Ayech
Department of Computer Science
University of Sherbrooke
J1K 2R1, Sherbrooke, Québec, Canada
ayechwalid@yahoo.fr

Djemel Ziou
Department of Computer Science
University of Sherbrooke
J1K 2R1, Sherbrooke, Québec, Canada
djemel.ziou@usherbrooke.ca

Abstract

Terahertz (THz) imaging is an innovative technology of imaging which can supply a large amount of data unavailable through other sensors. However, the higher dimension of THz images can be a hurdle to their display, their analysis and their interpretation. In this study, we propose a weighted feature space and a simple random sampling in k-means clustering for THz image segmentation. Our approach consists to estimate the expected centers, select the relevant features and their scores, and classify the observed pixels of THz images. It is more appropriate for achieving the best compactness inside clusters, the best discrimination of features, and the best tradeoff between the clustering accuracy and the low computational cost. Our approach of segmentation is evaluated by measuring performances and appraised by a comparison with some related works.

1. Introduction

Terahertz radiation (T-ray) refers to the region of electromagnetic spectrum occupying the band of frequencies from 0.1 to 10 THz and bounded by microwave and infrared bands. Compared to X-ray, infrared and microwave, THz image automatic analysis and interpretation are in its infancy. However, advances in THz acquisition technologies open the door to practical use in several areas such as medical diagnosis and security. The reader can find more about applications in [9, 11]. T-rays are non invasive and penetrate dry and non-metallic objects (paper, cloth, etc).

THz images can be acquired by acquisition in both active or passive modes. THz imaging in the active mode is formed by measures of sequences of chronological series or signals reflected from or transmitted through a sample. Each series can be represented by several bands or features (e.g. 1024 features) which characterise one pixel and the combination of these series into rows and columns constitutes the raw THz data cube (e.g. the $R \times C \times P$ cube in figure 1, where R, C and P represent respectively the number of rows, columns and features). Beyond the acquisition, the THz image segmentation has been studied in [8, 12, 13, 14, 15]. The k-means based clustering is the most popular used technique [1, 2, 5]. However, the high dimensionality of THz images lead to some new challenges for relevant feature selection. Indeed, the relevant features can be embedded only on few bands [4, 9]. For this raison, in several related works, some measures from the whole time series are used, such as the amplitude of the maximal pick and the time delay of the maximal pick of the time series [4, 8, 9]. These measures remain insufficient to characterise the different objects of THz images [1]. However, existing k-means algorithms deal all features with equal weights. Moreover, making use of the whole data into the clustering process decreases the k-means performances. It is thus necessary to integrate *data sampling* designs and *feature weighting* methods into k-means algorithm to extract relevant bands from the vast THz data observations. Data sampling and feature weighting techniques can improve the efficiency and the accuracy of the analysis.

In this paper we propose the use of SRS sampling and feature weighting. SRS sampling has been used in the works of Ayech and Ziou [2]. They have proposed an approach called SRS-k-means which consists to use the k-means technique under the SRS sample to avoid the use of the whole set of pixels. In this paper, we propose to reformulate the SRS-k-means by selecting the relevant features using a weighting strategy. The weighting methods have been used in [10] to analysis heart diseases and Australian credit card data. They have proposed to assign weights to features iteratively updated during the clustering process. Our main contributions consist to integrate the feature weighting and the SRS sampling into the k-means clustering. Our approach is called SS-k-means and consists to learn the weights of features according to their clustering importance and introduce the simple random sampling design into the k-means process. Note that the SS-k-means would be more appropriate for achieving the best compact-

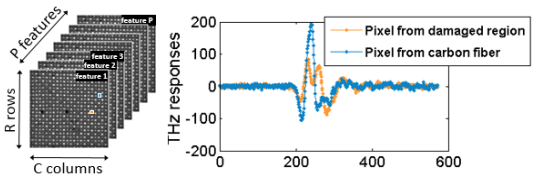


Figure 1. In the left, 3D THz data cube represented by $R \times C$ pixels and characterized by P raw features. The two pixels colored in blue and orange, which belong respectively to a typical region of the carbon fiber and a damaged zone, correspond to the two THz responses with the same colors in the right.

ness inside clusters, the best discrimination of features, and the best tradeoff between the clustering accuracy and the low computational cost. There are three main differences with the state of art: 1) the high dimensional data which are the THz images (1024 bands) and the feature space used; 2) the feature weighting formulation; 3) the combination of the feature weighting and the SRS sampling.

The paper is organized as follows: in section 2, we present an insight about related works of various THz imaging applications. Moreover, SRS-k-means approach has been detailed. We show its limitations and propose subsequently, in section 3, a novel approach to overcome these limitations. Our approach of segmentation is compared to k-means, SRS-k-means and W-k-means on THz images. The results are illustrated and discussed in section 4.

2. Background

2.1. Terahertz imaging

THz images are formed by capturing T-rays reflected from or transmitted through objects. Water and moisture objects highly absorb the T-rays, however, dry objects (such as paper, cloth and plastic) are transparent to T-rays and provide no significant reflected radiations. Metals are opaque to T-rays and reflect most incoming radiations. Other interesting materials, which offer specific T-rays, are detailed in [3, 9]. THz image is formed by several bands (e.g. 1024 bands). The high dimensionality of THz images leads to some new challenges for relevant feature selection. The features are used for the segmentation of THz images. In the most related works, classification of features is used for THz image segmentation.

Numerous works have been proposed to segment THz images. Some works are summarized in this section in terms of feature space used and classification or clustering techniques. The basic feature space is the raw time series of THz images [3]. Other feature spaces are obtained by using Fourier or Wavelet transforms [12, 13, 14]. The feature space can be used with only one band or with several bands. In the first case, the choice of the band can be priori fixed either from the time or the spectral spaces of the THz image [8]. Some measures from the shape of the entire time series or other spectral transform are used, such as the

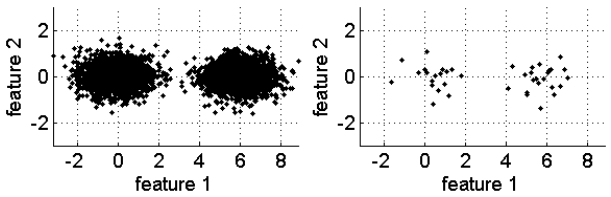


Figure 2. Dataset distributed in 2D space: full data ($N = 4000$) in the left and SRS sample ($n = 40$) in the right.

amplitude and the time delay of the maximal pick of time series [4, 8]. In the second case, several bands are used, such as the full time series of THz image, the full spectral amplitude, and a collection of some bands from time series [3, 4, 8, 13]. Some authors are proposed to reduce the feature space by using AR, ARMA, PCA, and decision tree [1, 6, 8, 14, 15]. THz image segmentation is generally performed in terms of classification, such as Mahalanobis classifier, SVM and neural networks [8, 13, 14, 15], and clustering, such as k-means, ISODATA, hierarchical chameleon, and KHM [1, 2, 3, 4, 6, 12].

In the most related works, k-means clustering has been shown efficient for the THz image segmentation [2, 3, 4, 6]. However, k-means techniques deal all the features with equal weights. Beside, making use of the whole data into the clustering process may decreases the k-means performances. We show in this paper the effect of both random sampling of the pixels and weighting method of the features in the k-means clustering accuracy. Our approach is compared with some related works.

2.2. SRS-k-means clustering

SRS-k-means [2] consists to combine the simple random sampling (SRS) [7] and the k-means clustering technique [5]. A representative sample X_{SRS} of n points from the observed population X of N points is then randomly selected and regrouped into homogeneous clusters in order to get conclusions about the centers. The main steps necessary to select a SRS data are summarized as follows:

SRS(X, n) *algorithm*

1. Develop a population list of all the elements of the studied population and assign each element a number to be able to access to the population.

2. Generate a list of n random numbers.

3. Select the elements $\{x_1, ..., x_n\}$ that have numbers corresponding to the generated random number list and save them in a dataset denoted X_{SRS}.

4. Return X_{SRS}.

Figure 2 shows an example of a dataset X (on the left) and a small X_{SRS} (on the right) distributed into two clusters. We show that only 1% of the population can represent the data under study and allows providing inferential statistics for the whole data clustering. SRS-k-means is a two-step algorithm. The E-step (estimation step) consists to classify the X_{SRS} into L clusters; each one is represented

by its own estimated center m_l. The C-step (clustering step) consists then to affect each point from the observed data X to the nearest cluster represented by its estimated center.

The SRS-k-means technique appears faster than the traditional k-means; however, all the features are dealt with equal importance into the clustering process in spite of some features can be noisy or uninformative.

3. The proposed SS-k-means clustering

In data clustering, there is no evident that features with equal importance will lead to the most significant clustering results. Traditional k-means clustering techniques deal with all features equally in deciding the cluster memberships of data points. However, this is not desirable in THz imaging where pixels often contains a huge number of diverse features. The structure of clusters in a given THz image is often restricted to a subset of features rather than the whole set of features. This leads us to ask the following questions: Is there a useful way to reduce the features space related to the structure of clusters? Is it possible to identify the relevant features for a given pixel? In this section, we start by presenting our feature weighting formulation into the clustering process to overcome the limitations of k-means techniques. Our approach, called SS-k-means, consists to combine the feature weighting and the simple random sampling into k-means clustering to provide the best tradeoff between the clustering accuracy and the low computational cost. The main idea of SS-k-means is to find a feature space in which the clusters are better separated. In other words, each cluster must possess a minimal dispersion, while the global data must be characterized by maximal dispersion. More formally, the dispersion within clusters is defined by:

$$\varphi_p = \sum_{l=1}^{L} \sum_{j=1}^{n} u_{jl}^a d(x_{jp}, m_{lp}) \tag{1}$$

where n represents the size of the SRS sample; m_{lp} is the center of the l^{th} cluster for the p^{th} feature; u_{jl} is the membership degree of the j^{th} point in the l^{th} cluster; $d(x_{jp}, m_{lp}) = (x_{jp} - m_{lp})^2$ is the distance metric that measures the similarity between a data point and a cluster center for the p^{th} feature, and $a > 1$ is the fuzzification degree. This criterion must be minimized to promote features having the best compactness inside clusters. The second criterion is called *global-data dispersion* criterion, represented by ψ_p and defined as follows:

$$\psi_p = \sum_{j=1}^{n} d(x_{jp}, m_p) \tag{2}$$

where m_p is the arithmetic mean of the SRS sample for the p^{th} feature. This criterion must be maximized to identify discriminative features which encourage the centers to be separated as much as possible.

Let us consider $w = (w_1, ..., w_P)$ be the weights for the P features and b a control parameter of these weights. A

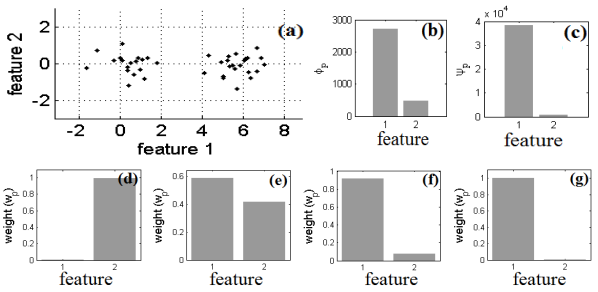


Figure 3. (a) X_{SRS} of $n = 40$ points. The SS-k-means clustering of X_{SRS} gives the within-cluster dispersion φ_p and the global-data dispersion ψ_p shown respectively in (b) and (c). (d), (e), (f) and (g) represent the final feature weights for $c = 0$, 0.5, 1 and 2.

compromise between minimizing φ_p and maximizing ψ_p leads to propose minimizing the following function:

$$J(\Phi, \Psi, w) = \sum_{p=1}^{P} w_p^b \frac{\varphi_p}{\psi_p^c} \tag{3}$$

where $\Phi = (\varphi_1, \cdots, \varphi_P)$ and $\Psi = (\psi_1, \cdots, \psi_P)$ are two vectors of P variables. The parameter c is a real which consists to control the effect of ψ_p regarding to φ_p. The objective function J is minimized with respect to the membership functions u_{jl}, the centers m_{lp} and the feature weights w_p. The variables u_{jl} and w_p must verify the constraints $\{u_{jl} \mid u_{jl} \in [0,1] \text{ and } \sum_{l=1}^{L} u_{jl} = 1\}$ and $\{w_p \mid w_p \in [0,1] \text{ and } \sum_{p=1}^{P} w_p = 1\}$. The objective function optimization is solved by using Lagrange multiplier technique. Then, the first order condition allows writing the membership degrees, the centers and the feature weights as follows:

$$u_{jl} = \left(\sum_{h=1}^{L} \left(\frac{\sum_{p=1}^{P} \frac{w_p^b}{\psi_p^c} d(x_{jp}, m_{lp})}{\sum_{p=1}^{P} \frac{w_p^b}{\psi_p^c} d(x_{jp}, m_{hp})} \right)^{\frac{1}{a-1}} \right)^{-1} \tag{4}$$

$$m_{lp} = \frac{\sum_{j=1}^{n} u_{jl}^a x_{jp}}{\sum_{j=1}^{n} u_{jl}^a} \tag{5}$$

$$w_p = \frac{\left(\frac{\psi_p^c}{\varphi_p}\right)^{1/(b-1)}}{\sum_{t=1}^{P} \left(\frac{\psi_t^c}{\varphi_t}\right)^{1/(b-1)}} \tag{6}$$

A representative sample from the observed population is then randomly selected in order to learn the cluster centers and the feature weights. We assume that the number L of clusters is known. The L-step (learning step) of SS-k-means consists to classify the X_{SRS} into L clusters of pixels; each cluster is represented by one center m_l and each pixel is characterized by P features and their weights w. The learning process is then done by iterating between three steps, updating the centers of clusters, the membership of objects and the weights of features, until convergence. Let us consider a parameter $Q \leq P$. The Q highest scores w^* are

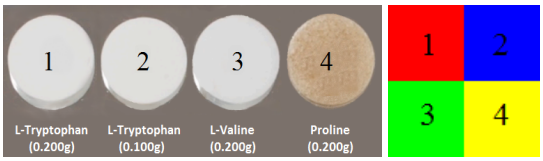


Figure 4. An image of four chemical compounds in visible spectrum in the left and the ground truth of the THz image in the right.

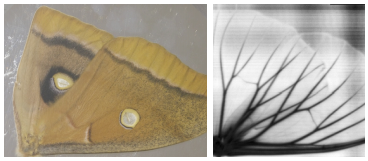


Figure 5. An image of a moth acquired in visible spectrum in the left and the 570^{th} band of the THz image in the right.

identified, the corresponding features are selected, and the dimensionality of the whole set of pixels are then reduced. The L-step algorithm can be summarized as follows:

L-step algorithm

1. Data: X_{SRS} = SRS(X,n) algorithm
Initialize m_l by random points from X_{SRS}

2. Do
Update centers m_{lq} using Eq. 5
Update membership degrees u_{jl} using Eq. 4
Update feature weights w_q using Eq 6
Until $|J^t - J^{t-1}|$ <threshold

3. Identify the Q highest weights w_p and select the corresponding features. Let us denote w^*, the vector of the selected feature weights.

The C-step (clustering step) of SS-k-means consists therefore to assign each point from the whole observed population, described in the space of the selected features, to the nearest cluster, represented by its estimated center m_l^*. For that reason, we propose to estimate the membership degree of data points by minimizing the objective function $J(\Phi^*, \Psi^*, w^*)$ in equation 3 where w^* is estimated in the L-step. The functions Φ^* and Ψ^* are described in the space of the selected features associated to the whole observed population. The membership degrees of the observed data to the clusters are given by equation 4 using Q, w^*, ψ_q^* and m^* instead of P, w, ψ_q and m. The resulted clusters are defined by the obtained membership degrees of data points.

Figure 3 (a) shows an example of X_{SRS} sample distributed in two clusters and randomly drown from the population X. The X_{SRS} sample represents only 1% of the observed population X in figure 2. We propose to cluster the population X by using the SS-k-means clustering. Figures 3 (b) and (c) show respectively the resulted φ_p and ψ_p associated to X_{SRS}. These figures show that $\varphi_1 > \varphi_2$ and $\psi_1 \gg \psi_2$. SS-k-means consists to promote features having a compromise between minimal values of within-cluster dispersion which corresponds to φ_2 and maximal values of global-data dispersion which corresponds to ψ_1. Figure 3 from (d) to (g) show the final feature weights (w_1 for feature 1 and w_2 for feature 2) respectively for c equal to 0, 0.2, 1 and 2. For c equal to 0, $w_2 > w_1$, while for 0.2, 1 and 2, $w_1 > w_2$ and w_1 grow when c increase. When $c > 0$, SS-k-means promotes then feature 1 than feature 2 which well explains the visual repartition of data. The example shows the interest of assigning weights to features by using a compromise between within-cluster and global-data dispersions associate only to a small number of data points. We note that k-means [5], W-k-means [10] and SRS-k-means [2] are particular cases of the SS-k-means. Indeed, The SRS-k-means can be obtained from equation 3 when $b = c = 0$, the W-k-means when $n = N$ and $c = 0$, and the k-means when $n = N$ and $b = c = 0$.

4. Experimental results

In this section, SS-k-means, W-k-means, SRS-k-means and k-means are tested on chemical and moth THz images. Since the THz images cannot be displayed (hundreds of bands), we present in figures 4 and 5 the objects which were acquired in THz spectrum and used for the validation. The ground truth of the chemical THz image and the 570^{th} THz band of the moth are shown in the right of the same figures. The chemical THz image is constituted by four compounds, L-Tryptophan (0.200g), L-Tryptophan (0.100g), L-Valine (0.200g) and Proline (0.200g), extracted and distributed into four false colored regions, while the moth image is essentially constituted by two wings. Each pixel of chemical and moth THz images are formed respectively by 1052 and 894 bands in the time domain. We have tested parameters a and b and we have found that best segmentation results are obtained when $a = b = 2$. The feature weights and the centers were randomly initialized by the same values for the different tests. The segmentation of the two images was employed respectively with 4 and 5 clusters. Figures 6 and 8 shows the chemical THz image segmentation for different techniques. In figure 6, the SS-k-means was carried out for $n = N$ using different values of c. For chemical compounds, both k-means and W-k-means produce as output an over-segmented images (figures 6 (a) and 6 (b)). In the case of k-means, L-Tryptophan (0.200g) and L-Tryptophan (0.100g) clusters are fused together which clearly shows the limitations of k-means segmentation using equal feature weights. However, W-k-means produces the final feature weights represented by the curve in figure 7 (a). The W-k-means promotes features in the interval [1,200] which are not discriminative and leads to over-segmented regions. Figure 6 from (c) to (f) display the obtained regions of SS-k-means for c equal to 0.5, 1, 1.2, and 1.5. For $c = 0.5$, the red region is ameliorated compared to W-k-means and begins to be clearly formed. The final feature weights are represented in figure 7 (b) and the highest scores is in the

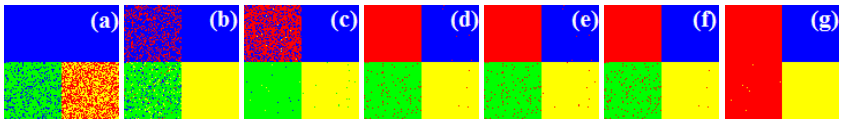


Figure 6. Chemical THz image segmentation for k-means (a), W-k-means (b) and SS-k-means for c=0.5 (c), 1 (d), 1.2 (e), 1.5 (f), 2.5 (g)

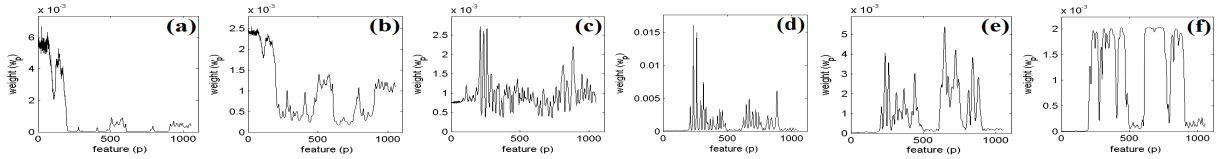
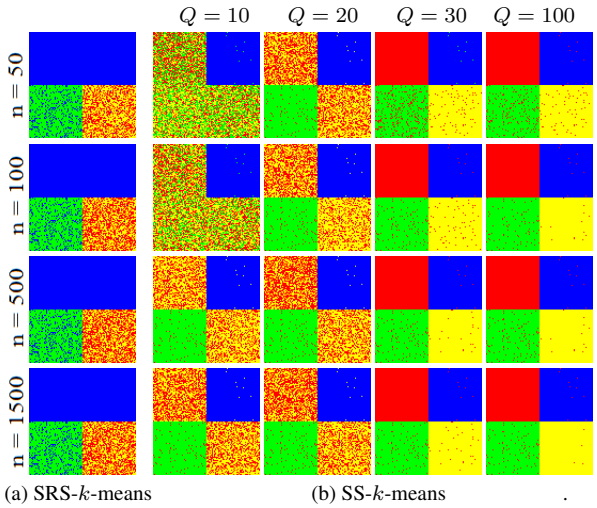


Figure 7. The feature weights of W-k-means (a) and SS-k-means on the chemical THz image for c=0.5 (b), 1 (c), 1.2 (d), 1.5(e) and 2.5(f)



(a) SRS-k-means (b) SS-k-means .

Figure 8. The chemical THz image segmentation of (a) SRS-k-means and (b) SS-k-means.

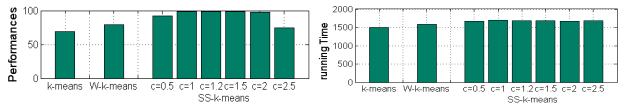


Figure 9. Clustering performance and run time of k-means, W-k-means and SS-k-means for n=N and c=0.5, 1, 1.2, 1.5, 2 and 2.5.

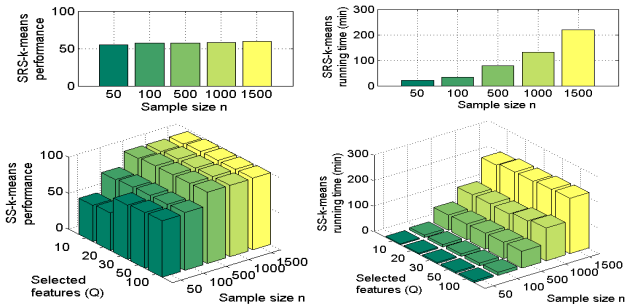


Figure 10. Clustering performances and running time of SRS-k-means and SS-k-means for different values of n and Q.

intervals [1,200], [490,680], and [910,1052]. Among them, some features are not yet discriminative to improve the clustering. The best chemical image segmentation is obtained when c surpass 0.5 in figures 6 (d), (e) and (f), the four compounds become very well identified, except some points of L-Valine (0.200g) are misclassified. The pertinent bands are around 250, 425, 610 and 720. However, when $c \geq 2.5$, SS-k-means segmentation of the chemical THz image produces under-segmented regions and the red and the green regions which represent respectively the L-Tryptophan (0.200g) and the L-Valine (0.200g) are fused together.

In figure 8, the chemical THz image segmentation using SS-k-means is also compared with the SRS-k-means for different sample size. The regions obtained by SRS-k-means are very bad and similar to k-means output. However for $c = 1.2$, the figure 8 (b) shows the output regions of SS-k-means for different values of n and Q. Note that when Q surpassing 20, the results are very interesting. For a good choice of n and Q, the statistics in figures 9 and 10 show the efficiency and the rapidity of SS-k-means to segment the chemical THz image. Then a small sample of pixels, around 1%, and a minimal number of features, around 3%, are sufficient to produce good segmentation using SS-k-means.

Figures 11 and 13 show the segmentation outputs of the four clustering algorithms on the moth THz images. Both k-means and W-k-means produce a wrongly segmented regions in figures 11 (a) and 11 (b). The obtained regions clearly illustrate the limitations of both techniques to provide good structure of wings. Figure 12 (a) shows that the weights estimated by using W-k-means in the intervals [1,100] and [220,380] are not relevant, which leads to the under-segmentation. Figure 11 from (c) to (f) display the obtained regions of SS-k-means for c equal to 1, 1.5, 2, and 2.5. The structure of the moth wings is again wrongly segmented for c equal to 1 and 1.5. The corresponding highest feature weights in figure 12 (b) and (c) are around 150, 400, and 680. Among them, some features are not relevant which explain the decrease of the clustering performances. The best regions are obtained when c surpass 1.5 in figures 11 (e) and (f). The structure of the moth wings are preserved. In figure 13, a comparison between SS-k-means and SRS-k-means are done. The regions output of SS-k-means (c fixed to 2) outperform those obtained by SRS-k-means for different values of n and Q. Note that only less than 1% and low number of features are sufficient to segment accurately both images.

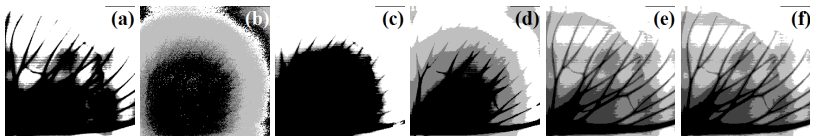


Figure 11. Moth THz image segmentation for k-means (a), W-k-means (b) and SS-k-means for $c = 1$ (c), 1.5 (d), 2 (e), 2.5 (f)

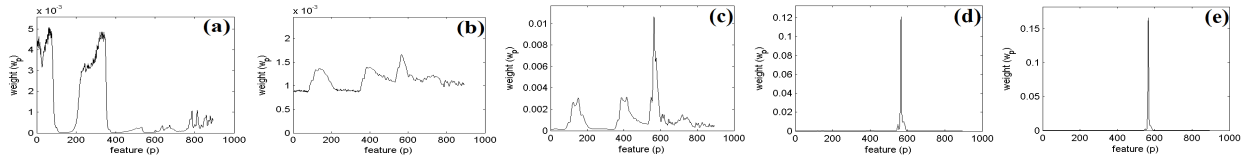
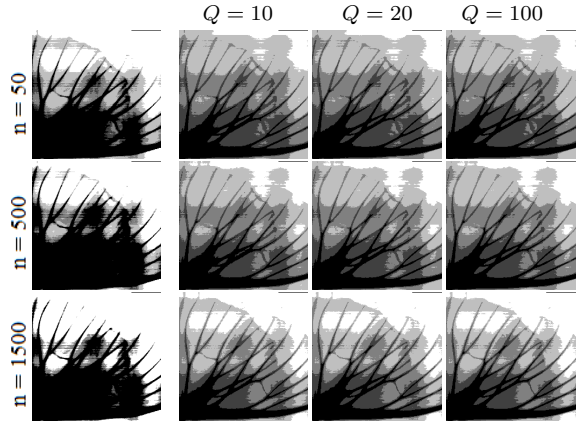


Figure 12. The feature weights of W-k-means (a) and SS-k-means on the moth THz image for c = 1 (b), 1.5 (c), 2 (d) and 2.5 (e)



(a) SRS-k-means (b) SS-k-means .

Figure 13. The moth THz image segmentation of (a) SRS-k-means and (b) SS-k-means.

5. Conclusion

In this paper, we have proposed a novel clustering approach, called SS-k-means, to segment THz images. Feature weighting is used in order to reduce the number of features required for carrying out the segmentation. In addition to the computational time, irrelevant features decreases the clustering accuracy. The SRS scheme allows to use around 1% of pixels. Our approach is more appropriate for achieving the best compactness inside clusters and the best discrimination of features. It is evaluated and compared favorably with some related works. Note that, the choice of the SRS sample size and the number of features are important issues which require further studies.

Acknowledgments

The authors would like to thank Prof. Thomas Tongue of Zomega THz Corporation for the THz measurements.

References

[1] M. W. Ayech and D. Ziou. Terahertz image segmentation based on k-harmonic-means clustering and statistical feature extraction modeling. In *ICPR*, pages 222–225, Tsukuba, Japan, 2012.

[2] M. W. Ayech and D. Ziou. Segmentation of terahertz imaging using k-means clustering based on ranked set sampling. *Expert Systems with Applications*, 42(6):2959 – 2974, 2015.

[3] E. Berry, R. D. Boyle, A. J. Fitzgerald, and J. W. Handley. Time frequency analysis in terahertz pulsed imaging. In I. Pavlidis, editor, *CVBVS*, Advances in Pattern Recognition, chapter 9, pages 271–311. Springer Verlag, 2005.

[4] E. Berry, J. W. Handley, A. J. Fitzgerald, W. Merchant, R. D. Boyle, N. Zinovev, R. E. Miles, J. Chamberlain, and M. A. Smith. Multispectral classification techniques for terahertz pulsed imaging: an example in histopathology. *Med. Eng. & Phys.*, 26(5):423–430, 2004.

[5] J. C. Bezdek. *Pattern recognition with fuzzy objective function algorithms*. Plenum, New York, NY, 1981.

[6] M. A. Brun, F. Formanek, A. Yasuda, M. Sekine, N. Ando, and Y. Eishii. Terahertz imaging applied to cancer diagnosis. *Phys. in Med. and Bio.*, 55(16):4615–4623, 2010.

[7] W. G. Cochran. *Sampling Techniques*. John Wiley & Sons, New York, third edition, 1977.

[8] L. H. Eadie, C. B. Reid, A. J. Fitzgerald, and V. P. Wallace. Optimizing multi-dimensional terahertz imaging analysis for colon cancer diagnosis. *ESWA*, 40(6):2043–2050, 2013.

[9] A. J. Fitzgerald, E. Berry, N. N. Zinovev, G. C. Walker, M. A. Smith, and J. M. Chamberlain. An introduction to medical imaging with coherent terahertz frequency radiation. *Phys. in med. & bio.*, 47(7):R67–R84, 2002.

[10] J. Z. Huang, M. K. Ng, H. Rong, and Z. Li. Automated variable weighting in k-means type clustering. *IEEE TPAMI*, 27(5):657–668, 2005.

[11] D. M. Mittleman, M. Gupta, R. Neelamani, R. G. Baraniuk, J. V. Rudd, and M. Koch. Recent advances in terahertz imaging. *App. Phys. B: Lasers and Optics*, 68:1085–1094, 1999.

[12] H. Stephani. *Automatic Segmentation and Clustering of Spectral Terahertz Data*. PhD thesis, 2012.

[13] X. Yin, B. W. H. Ng, B. M. Fischer, B. Ferguson, and D. Abbott. Support vector machine applications in terahertz pulsed signals feature sets. *Sens. Jour.*, 7(12):1597–1608, 2007.

[14] X. X. Yin, B. W. H. Ng, B. Ferguson, and D. Abbott. Statistical model for the classification of the wavelet transforms of t-ray pulses. In *ICPR*, volume 3, pages 236–239, 2006.

[15] H. Zhong, A. Redo-Sanchez, and X. C. Zhang. Identification and classification of chemicals using terahertz reflective spectroscopic focal-plane imaging system. *Optics Express*, 14(20):9130–9141, Oct 2006.