Expert Systems with Applications 42 (2015) 2959–2974

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Segmentation of Terahertz imaging using *k*-means clustering based on ranked set sampling

Mohamed Walid Ayech *, Djemel Ziou

Department of Computer Science, University of Sherbrooke, J1K 2R1 Sherbrooke, Québec, Canada

ARTICLE INFO

Article history:
Available online 29 November 2014

Keywords:
Segmentation
Terahertz imaging
k-means
Ranked set sampling
Simple random sampling

ABSTRACT

Terahertz imaging is a novel imaging modality that has been used with great potential in many applications. Due to its specific properties, the segmentation of this type of images makes possible the discrimination of diverse regions within a sample. Among many segmentation methods, *k*-means clustering is considered as one of the most popular techniques. However, it is known that *k*-means is especially sensitive to initial starting centers. In this paper, we propose an original version of *k*-means for the segmentation of Terahertz images, called ranked-*k*-means, which is essentially less sensitive to the initialization of the centers. We present the ranked set sampling design and explain how to reformulate the *k*-means technique under the ranked sample to estimate the expected centers as well as the clustering of the observed data. Our clustering approach is tested on various real Terahertz images. Experimental results show that *k*-means clustering based on ranked set sampling is more efficient than other clustering techniques such as the *k*-means based on the fundamental sampling design simple random sampling technique, the standard *k*-means and the *k*-means based on the Bradley refinement of initial centers.

1. Introduction

In recent years, many research groups around the world are increasing their interest on the Terahertz (THz) portion of the electromagnetic radiation (Hu & Nuss, 1995; Zhang, 2002). Terahertz radiations (T-rays) have been used in many applications, due to their interesting properties, such as noninvasive property, penetration through dry and non-metallic objects (plastic, paper, cloth, etc), and specific material characterization. The use of T-rays for imaging has opened new possibilities for research and commercial applications (Bowen et al., 2013; Fischer, Hoffmann, Helm, Modjesch, & Jepsen, 2005; Fitzgerald et al., 2002; Ikushima & Komiyama, 2010; Kamba, Tsuchiya, Okimoto, & Fukunaga, 2011; Kowalski, Palka, Piszczek, & Szustakowski, 2013; Mittleman et al., 1999; Mittleman, Jacobsen, & Nuss, 1996; Willis & Wilson, 2013).

Terahertz pulsed imaging (TPI) system consists in collecting information from the scene, as a sequence of two-dimensional images. Each image is constituted by a set of grey level pixels acquired from a single spectral band. The combination of these images constitutes a three-dimensional Terahertz data cube. Compared to the color imaging, each pixel in the Terahertz imaging acquires many bands (e.g. 1024 bands) from the electromagnetic spectrum, instead of the only three bands of the RGB color representation. TPI system can provide specific temporal and spectral information unavailable through other sensors characterizing each pixel of the THz image. The segmentation of THz imaging supplies a wealth of information about test samples and makes possible the discrimination of heterogeneous regions within an object. Among many segmentation methods, *k*-means clustering (Ayech & Ziou, 2012; Bezdek, 1981; Jain, Murty, & Flynn, 1999; MacQueen, 1967) is considered as one of the most popular techniques developed in the last few decades, due to its simplicity of implementation, fast execution and good computational performance. However, it is well known that *k*-means might converge to one of numerous local minima, and its result depends on initial starting conditions, which randomly generates the initial clustering (Jain et al., 1999). In other words, different clustering results can be produced after different runs of *k*-means on the same input data. Given an association rule between the data points and the centers, the clustering accuracy depends on the location of the centers. The structure of the data and the sampling procedure has an effective impact on the estimation of the centers. In machine learning, the impact of sampling is often unmentioned. We show in this paper the effect of the sampling procedures in the clustering process. Simple random sampling (SRS) is the mostly used procedure in

* Corresponding author.
E-mail addresses: ayechwalid@yahoo.fr (M.W. Ayech), djemel.ziou@usherbrooke.ca (D. Ziou).

http://dx.doi.org/10.1016/j.eswa.2014.11.050
0957-4174/

which the data points are assumed to be iid (Cochran, 1977; Thompsom, 1997) and there are only a few results available when the sampling design is different (Bejarano et al., 2011; Bradley & Fayyad, 1998; Chen, 2010). However, in some applications, such as the one explained in MacQueen (1967), Patil (2002) and Patil, Sinha, and Taillie (1994), using ranked set sampling (RSS), may be cheaper and result in better and more informative samples from the underlying population. In this paper, we study the problem of initial center sensitivity of k-means technique; explain how to reformulate the k-means under the RSS design to overcome the initialization problem and classify the observed data. The obtained results are compared with the corresponding ones of simple random sample data. We show that, using RSS, our approach leads to a better inference about the precision of centers and therefore the precision of the obtained clusters. Experimental tests of our approach are done to segment Terahertz images. The obtained results show the interest of ranking the pixels and explain how the extra information via the rank of each pixel in RSS will lead to a more efficient classification of pixels compared with SRS and other techniques.

The rest of the paper is organized as follows: in Section 2, we give an insight about related works of various imaging applications in the Terahertz domain. Section 3 introduces the k-means clustering based on the simplest sampling design SRS technique that we call SRS-k-means. Section 4 presents the RSS technique and explains its efficiency compared to the SRS. The formulation of the general k-means in the case of RSS sample and the different steps of the resulting algorithm, ranked-k-means, are also described. Our clustering approach based on RSS sample is compared with the clustering approach in the case of SRS, the standard approach of k-means and the k-means using the Bradley refinement of initial centers on the real Terahertz images of a carbon fiber sample, a flexure spring and a fruit grape. The results are illustrated and discussed in Section 5.

2. Related works on Terahertz image segmentation

The Terahertz image is formed by capturing THz radiations reflected from or transmitted through objects. Water and moisture objects highly absorb THz radiations, however, dry objects (such as paper, cloth, plastic and wood) are transparent to THz radiations and provide no significant reflected radiations. Metals are opaque to THz radiations and reflect most incoming radiations. Other interesting materials, which offer specific THz radiations, are detailed in Fitzgerald et al. (2002) and Berry, Boyle, Fitzgerald, and Handley (2005). The 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 detection. Indeed, the relevant features can be embedded only on few bands (Berry et al., 2005; Fitzgerald et al., 2002). For this raison, in several related works, the band having the maximal pick amplitude is used. It has pointed out that other bands may contains relevant features and these bands are not known in advance (Ayech & Ziou, 2012; Nakajima, Hoshina, Yamashita, Otani, & Miyoshi, 2007; Yin, Ng, Ferguson, Mickan, & Abbott, 2006). The features are used for the segmentation of THz images. In the most related works, classification of features is used for the segmentation of Terahertz images.

Table 1 presents a summary of several segmentation methods, regrouped in terms of feature space used, classification or clustering techniques and application domains. From this table, we deduce three important remarks. The first one concerns the various application domains using the Terahertz imaging which explains the interest of analysing this now technology of imaging. The second remark concerns the feature spaces used in the state of

Table 1
Summary of some related works on Terahertz image segmentation.

Methods	Features	Classification & clustering	Application domains
Berry et al. (2004)	TD & MxA & FWHM	k-Means & ISODATA	Histopathology (basal cell carcinoma and melanoma diagnosis)
Berry et al. (2005)	Time series & Short time Fourier transform & DWT	k-Means	Medical (dental) & histopathology(carcinoma and melanoma diagnosis)
Yin et al. (2006)	DWT & AR/ARMA	Mahalanobis distance classifier	Biomedical (osteosarcoma cells diagnosis) & security (mail/ packaging inspection)
Zhong et al. (2006)	PCA of the interval 0.4–1.6 THz	Minimum distance classifier & NN	Chemistry (material identification)
Yin et al. (2007)	Spectral magnitude & spectral phase	SVM	Biomedical (ribonucleic acid recognition) & chemistry (powder identification)
Nakajima et al. (2007)	PCA	k-Means & AH	Histopathology
Brun et al. (2010)	PCA	fuzzy k-means	Histopathology (cancer inspection from lung and pancreas)
Stephani (2012)	DWT	Hierarchical chameleon	Security (mockup mail bomb detection)
Ayech and Ziou (2012)	PCA & AR	KHM	Quality control (damage tasks detection) & agricultural (crop yield estimation)
Eadie et al. (2013)	MxA, MnA, -MnA/ MxA, MxA-MnA, FWHM, T(t), F(f), T(t)/MnA, & decision tree	NN & SVM	Medical (colon cancer diagnosis)