








Automatic Cluster Selection in K-Means Lung Segmentation

Natanael Hernández-Vázquez^(✉) , Stewart René Santos-Arce ,
Ricardo Antonio Salido-Ruiz , Sulema Torres-Ramos ,
and Israel Román-Godínez 

Division of Cyber-Human Integration Technologies, University of Guadalajara,
44430 Guadalajara, Jalisco, Mexico
natanael.hernandez4729@alumnos.udg.mx

Abstract. Lung segmentation is a critical step in machine-learning-based radiomics using thoracic computed images. It involves isolating a specific region of interest, but variations in lung intensity values caused by diseased lung tissue can difficult correct segmentation. Although K-means is commonly used, it requires manual intervention to select each cluster related to the region of interest, leading to an efficiency decrease in terms of the specialist's time and effort, especially for large image volumes. To address these limitations, an automatic cluster selection methodology is proposed. It involves a training process to determine a threshold for discriminate clusters; then, morphological transformations and image processing techniques enhance segmentation. Evaluation using DICOM images from the Interstitial Lung Diseases Database yielded a Jaccard Similarity Index of 0.9056 and a Dice Similarity Coefficient of 0.9475, demonstrating the effectiveness and accuracy of the proposed approach.

Keywords: Lung segmentation · K-means · Computational Vision · Image Processing

1 Introduction

Segmentation is the process of delineating specific areas within an image, achieved by assigning labels to each pixel based on regions of interest (ROIs). Various methodologies exist for performing semi-automatic or automatic segmentation, intelligent image processing analysis in the scenario [1, 2]. Thresholding is an easiest algorithm used for image segmentation which groups pixels based on intensity differences [2].

In lung images, segmentation serves as a preprocessing step to separate the lungs, enabling various machine learning tools to concentrate their actions exclusively on this tissue, thus improving performance. Thresholding algorithms work well when lung pathologies are absent, as there is a noticeable contrast between the lungs and the surrounding tissue in both computed tomography (CT) images and plain X-rays. However, complications arise when the lung tissue density increases due to diseases such as pulmonary fibrosis [3], interstitial lung disease [4], and cancer [5], among others, resulting in X-ray beams interacting similarly with both lung and surrounding tissues [6, 7].

To overcome this challenge, state-of-the-art (SOA) approaches often employ K-means due to their simplicity and interpretability. For instance, Gupta et al. (2022) used K-means, fuzzy C-means to separate anatomical structures within the CT images, incorporating wavelet techniques to enhance the obtained mask. Their method achieved an accuracy, Dice Similarity Coefficient (DSC), and Jaccard Similarity Index (JSI) of 0.9928, 0.9872, and 0.9787, respectively [1]. Similarly, Hu et al. (2020) employed Convolutional Neural Networks for lung region mapping and utilized Bayes, Support Vector Machines, and K-means as kernels, achieving an accuracy of 0.97 [8]. Liu et al. (2023) utilized K-means in conjunction with Hough transform to remove cavities, obtaining a DSC and JSI of 0.9786 and 0.9512, respectively [9].

While these studies effectively tackle lung segmentation (LS), they rely on the manual identification of lung clusters per image to generate masks. This can be time-consuming for specialists dealing with sizable image volumes during diagnosis and treatment, where accuracy is vital. Hence this work proposes an approach to automate the selection of lung and non-lung clusters. The goal is to have a methodology for the segmentation of diseased lungs that achieves competitive performances regarding SOA methods.

2 Materials and Methods

K-means is an unsupervised clustering algorithm designed to identify K groups within a dataset. In the context of image segmentation, the dataset consists of elements represented by a dimension $M \times N$ image matrix denoted as \mathbf{X} . The algorithm groups the pixels in \mathbf{X} into K based on their similarity. The user determines the number of clusters, K , based on prior \mathbf{X} data analysis. The resulting segmentation, \mathbf{Y} , allows for drawing conclusions and characterizing the K groups [10].

2.1 Image Segmentation Process using K-means

The image segmentation process is showed in Fig. 1 [11]. It involves treating the image as a dimension $M \times N$ matrix, denoted as \mathbf{X} , where each element X_{mn} represents the intensity information. In CT images, this intensity is expressed in Hounsfield Units [12]. To facilitate processing, the matrix is transformed into a vector using lexicographical ordering $\mathcal{L}\{\cdot\}$, this consists of the column-by-column to left-to-right stacking of the matrix \mathbf{X} [13]. This vector is then subjected to the K-means algorithm, resulting in a vector where each pixel is assigned to one of the K clusters. Then K-means clustering model output is reorganized through lexicographical reordering $\mathcal{L}^{-1}\{\cdot\}$, resulting in a \mathbf{Y} matrix of dimension $M \times N$. With this, the user can identify and select the clusters of interest. In the context of LS, the user designates all the pixels within the lung clusters with an intensity value of 1, while non-lung pixels are assigned an intensity value of 0. This process is referred to as binarization, given its binary decision nature in this study.

One drawback of the image segmentation process using K-means is the manual assignment of labels (such as lung or not-lung) to each cluster in every image, which can compromise its effectiveness.

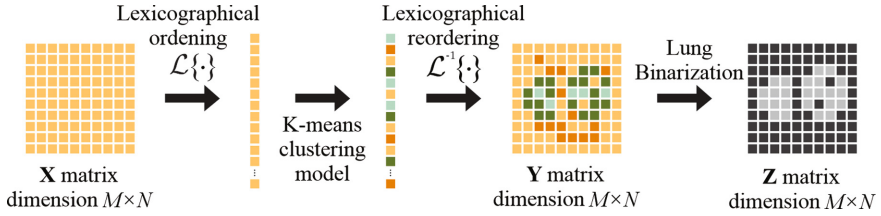


Fig. 1. Representation of image segmentation process using K-means.

2.2 Automatic Cluster Selection

To address the issue of manual selection and labeling of clusters in K-means LS, an automatic methodology is proposed (see Fig. 2). The main idea of this methodology is to determine an upper threshold α and a lower threshold β . Once these thresholds are computed, the ratio of pixels within each cluster that falls within the limits of α and β is calculated.

First, the data set A is divided into two sets: 70% of the studies for training (A_1) and 30% for test (A_2). The training set, A_1 , contains \mathbf{B}_i thorax CT images, while the test set, A_2 , consists of \mathbf{C}_i thorax CT images. Additionally, the \mathbf{D}_i mask is required as a counterpart for \mathbf{C}_i and \mathbf{B}_i ; every image matrix of $M \times N$ dimension. The \mathbf{D}_i mask contains labeled pixels indicating whether they belong to lung and non-lung regions, with lung pixels assigned an intensity value of 1 and non-lung pixels assigned an intensity value of 0. This resulting mask serves as gold-standard for each \mathbf{B}_i and \mathbf{C}_i images.

The training process involves the following steps: for each \mathbf{B}_i image in A_1 , the \mathbf{D}_i mask is applied to compute the element-wise product (\times) between \mathbf{B}_i and \mathbf{D}_i resulting in the extraction of lung intensities. This procedure is repeated for all A_1 images. Using the lung intensities from all A_1 images, the global mean \bar{x} and global standard deviation σ are calculated. These values are then used to determine the threshold using the following equations:

$$\alpha = \bar{x} + \sigma, \quad (1)$$

$$\beta = \bar{x} - \sigma. \quad (2)$$

During the experiment, all \mathbf{C}_i images from A_2 are collected and concatenated into a matrix \mathbf{C}' with dimensions $M \times N_i$. Subsequently, the K-means image segmentation process, as described in Sect. 2.1, is applied to \mathbf{C}' .

To determine the number of clusters, various experiments were conducted with values of k ranging from 2 to 10. Optimal results were achieved with 4 clusters. Through the K-means segmentation process every pixel in the matrix \mathbf{C}' is assigned a label corresponding to one of the K clusters. To determine whether a cluster represents the lungs, a ratio is calculated based on the pixels within the threshold defined by α and β . This ratio is computed using the following equation:

$$\gamma = \frac{p}{P} \quad (3)$$

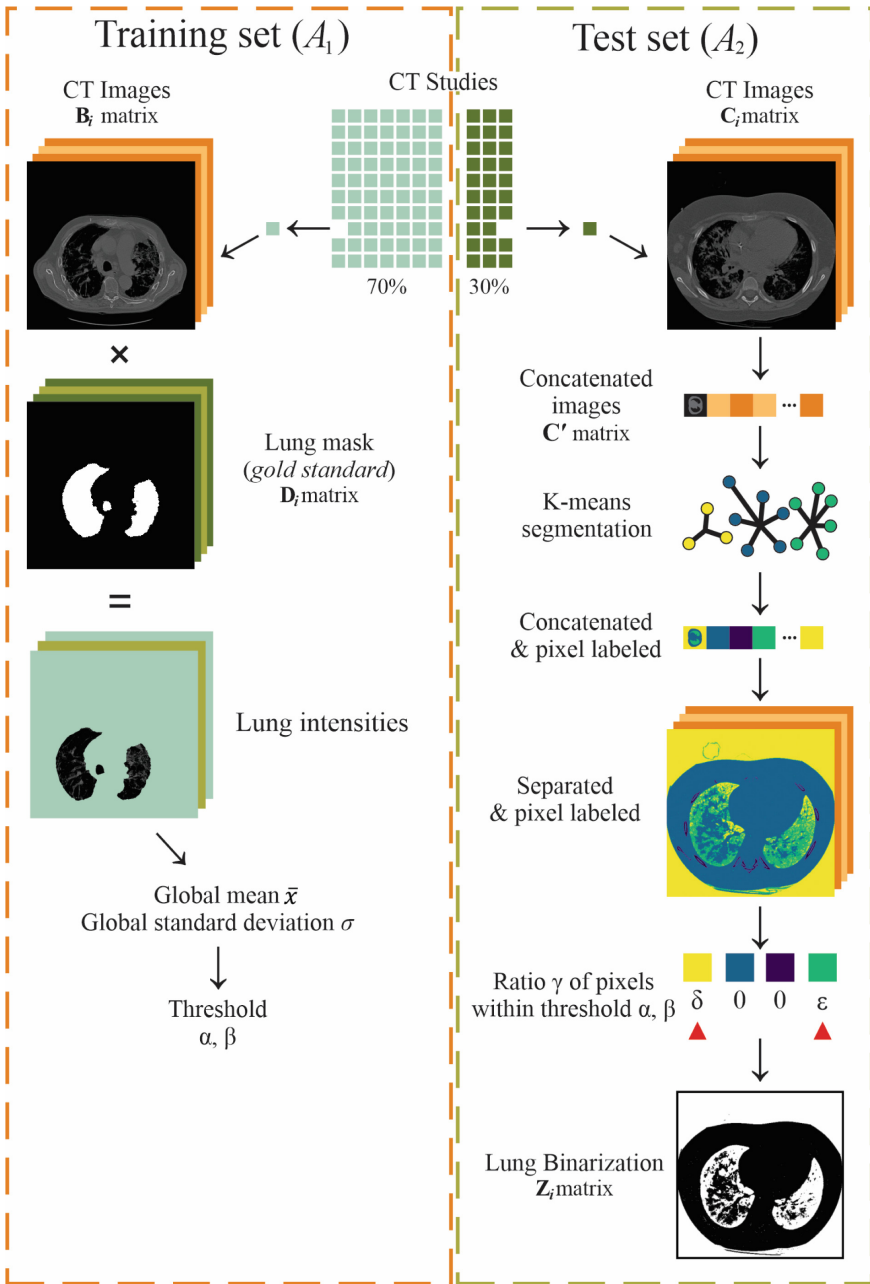


Fig. 2. Schematic of the proposed methodology. δ and ϵ are values greater than 0.001.

where p is the number of pixels within the threshold for a given cluster, while P represents the total number of pixels in that cluster. By calculating γ , clusters with values greater than 0.001 are selected. Subsequently, all pixels belonging to the selected clusters are assigned a value of 1, while the remaining pixels are assigned 0. This generates the $M \times N_i$ matrix binary mask C_i , that needs to be divided into i separate matrices. As a result, we obtain Z_i matrices with $M \times N$ dimensions.

2.3 Post-processing Mask Strategy

The resulting Z_i is obtained for each C_i and undergoes post-processing, which involves applying morphological transformation and image-processing techniques. This process aims to refine the mask and ensure that it covers most lung pixels. Further details of this post-processing procedure are shown in Fig. 3.

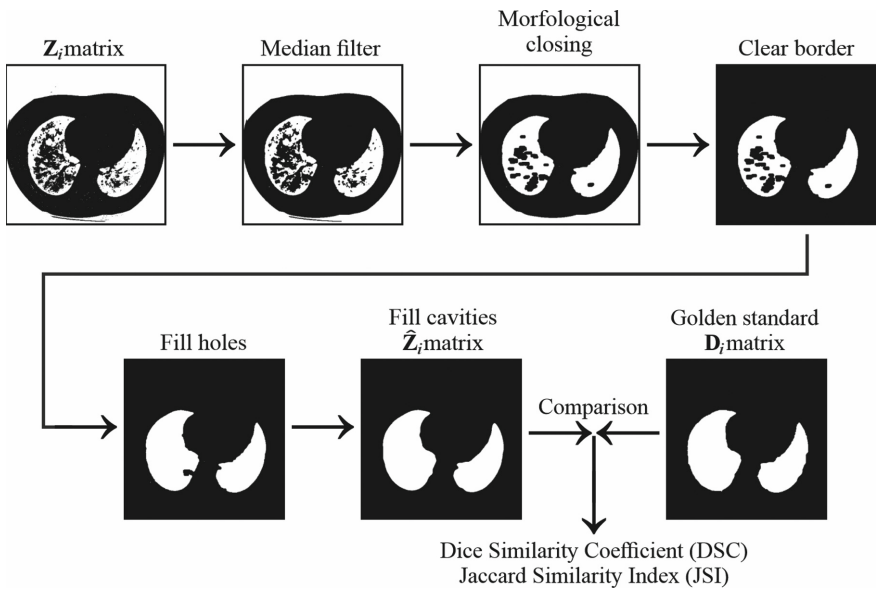


Fig. 3. Schematic post-processing including the gold-standard comparison.

The initial step involves applying a median filter as suggested by Tukey [13], with the defined kernel size of three pixels. This process eliminates scattered pixels in the image and smooth out the relevant ones. Next, a morphological closing operation, initially proposed by Matheron and Serra [14] is performed. The image undergoes dilation followed by an erosion to enhance the solidity of the mask. A circular structural element with a radius of eight pixels is utilized for this operation.

The presence of air outside the body and within the lungs leads to similar intensity values, making it necessary to remove it from the mask. To accomplish this, a morphological transformation is applied, taking advantage of the proximity of air pixels along the image edge. By intersecting the input image with its edge, a marker image is created,

containing seeds for each connected pixel or particle at the edge. Through reconstruction, an image consisting of these particles is obtained and subsequently erased [15]. To fill any remaining unconnected holes, an erosion-based reconstruction is performed using the mask and a marker image with consistent lung values [15].

Depending on the disease, cavities can appear at the edge of the region of interest (ROI). To address this, the circular shape of the cavities can be leveraged using an algorithm described by Liu et al. [9]. The algorithm utilizes the Hough transform to detect circles to be filled. Selection criteria were defined as follows: circles with less than 1/2 of lung pixels, and circles with more than 2/3 of lung pixels on their perimeter. This process results in the post-processing matrix \hat{Z}_i with $M \times N$ dimensions.

2.4 Evaluation Metrics

To assess the performance of the proposed methodology, the post-processed matrix \hat{Z}_i was compared to the corresponding gold-standard mask D_i using DSC [16] and JSI [17]. These metrics provide a similarity value ranging from 0 to 1, where 0 indicates no spatial overlap and 1 represents complete spatial agreement.

Following extensive comparisons using DSC and JSI metrics, the global mean and standard deviation for each metric were calculated and then compared to previous works in the SOA.

We use DSC and JSI metrics in this study based on their common usage in segmentation methodologies and their ability to evaluate performance. However, it is important to note that JSI offers advantages over DSC. Unlike DSC, JSI satisfies all the properties of a metric, including the crucial triangular inequality property [18]. The relaxed triangular inequality in DSC can affect efficiency and approximation ratios, rendering it not fully considered as a metric [19].

2.5 Computational Tools

This methodology and experimentation were developed using Python 3.9.16. Pydicom 2.3.1 for reading DICOM files and their metadata. OpenCV 4.7 for image morphological transformations, NumPy 1.21.5 and Pandas 1.5.3 for matrix analysis, operations, data transformation, and structures, and Matplotlib 3.6.3 for image and results displaying. All this work was carried out using the hardware CPU AMD Ryzen 7 5800 H 3.20 GHz, GPU Nvidia GeForce RTX 3060 6 GB VRAM and 16 GB RAM.

3 Results and Discussion

The evaluation of the proposed methodology use the Multimedia Database of Interstitial Lung Diseases created by Depeursinge et al., and its use for research purposes permitted by the ethics committee of the University Hospitals of Geneva [20]. This database comprises 3076 CT images in DICOM format, obtained from 113 patients diagnosed with various lung diseases within ILDs. Each image has dimensions of 512×512 pixels and is of uint16 value type.

Additionally, a gold-standard lung mask, manually annotated by a medical specialist, is available for each image in the database. For the quantitative assessment, the metrics DSC and JSI were employed. Figure 4 presents box plots illustrating the outcomes obtained by applying the proposed methodology to all 3076 images split randomly into a training set (70% - 2,153 images) and test set (30% - 923 images). Furthermore, the results reported by Gupta et al. [1] and Liu et al. [9] are also depicted in Fig. 4.

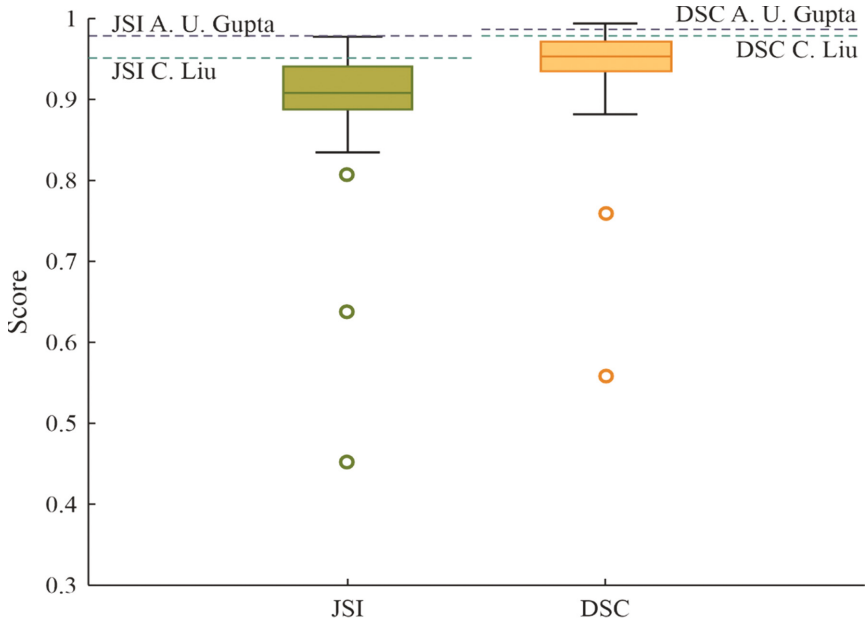


Fig. 4. Performance of the proposed methodology showed in JSI and DCS.

The performance of the proposed automatic cluster selection using K-means can be observed. The box plot reveals a low dispersion of 0.0660 and 0.0520 for JSI and DSC respectively. Notably, Quartile 1 for JSI is 0.887 and for DSC is 0.9351, while Quartile 3 for JSI is 0.9403, and for DSC is 0.9723. These results indicate that approximately half of the automatic segmentation indexes fall within these ranges, suggesting stability in the proposal. However, there are three outliers for JSI and two outliers for DSC, implying that certain images pose challenges for segmentation using this approach. Hence, it is expected to achieve performances close to the median values of 0.9092 for JSI and 0.9539 for DSC. In the comparison with the SOA, Fig. 4 demonstrates that the segmentation results obtained by this proposal are comparable to Gupta et al. [1], and in certain cases, they even surpass the results reported by both authors.

Additionally, Table 1 presents the results obtained using the same metrics and dataset, with K-means as the clustering method. First, Gupta's approach [1] incorporated fuzzy C-means and wavelets into their methodology, while Liu et al.'s approach [9] incorporated the Hough transform. Both authors noted that manual intervention is required for

selecting lung or not-lung clusters. Gupta et al. also reported the need for manual selection of clusters during image decomposition and reconstruction using wavelets, which further increases the level of user intervention required for creating the mask.

Table 1, demonstrates that with the proposed automatic cluster selection approach, it is possible to achieve performances above 0.90 for both JSI and DSC metrics, with a standard deviation of 0.066 or lower. In contrast, Gupta et al. [1] and Liu et al. [9] did not provide information on error rate results or standard deviation resulting from manual interaction. Consequently, the potential impact of this manual selection on reproducibility within their methodologies cannot be determined.

Table 1. Comparison of the proposal with the SOA.

Methodology	JSI		DSC	
	<i>Mean</i>	<i>Standard Deviation</i>	<i>Mean</i>	<i>Standard Deviation</i>
Gupta et al. [1]	0.9787	-	0.9872	-
Liu et al. [6]	0.9512	-	0.9786	-
Proposed methodology	0.9056	0.0660	0.9475	0.0520

Furthermore, results above 0.90 in DSC and JSI (see Table 1) were accomplished, which are near to the values reported by each author, but without the manual intervention of experts that spend important time and effort.

4 Conclusions

This work introduces a methodology that enables automatic cluster segmentation eliminating the need for manual selection of clusters by experts. This approach facilitates fully automated LS, which can be valuable for various applications such as disease classification and delineating ROIs within the lungs for radiological analysis allowing to decrease specialist's work and time spent in the analysis of large image volumes.

In future work, this methodology will be enhanced by incorporating additional image characteristics such as textures and pixel relationships. It is also crucial to evaluate the performance of the proposed methodology on different datasets to assess its robustness. By testing the methodology in diverse scenarios, its applicability and generalization capabilities can be examined.

References

1. Gupta, A.U., Singh-Bhadauria, S.: Multi level approach for segmentation of interstitial lung disease (ILD) patterns classification based on superpixel processing and fusion of k-means clusters: SPFKMC. *Comput. Intell. Neurosci.* **2022**, 22 (2022)

2. Pham, D.L., Xu, C., Prince, J.L.: Current methods in medical image segmentation. *Annu. Rev. Biomed. Eng.* **2**(1), 315–337 (2000)
3. Christie, A., Peters, A.A., Drakopoulos, D., Heverhagen, J.T., Geiser, T., Stathopoulou, T., et al.: Computer aided diagnosis of pulmonary fibrosis using deep learning and CT images. *Invest. Radiol.* **54**(10), 627–632 (2019)
4. Pawar, S.P., Talbar, S.N.: Two stage hybrid approach of deep learning networks for interstitial lung disease classification. *Biomed. Res. Int.* **2022**, 1–10 (2022)
5. Liu, X., Li, K.W., Yang, R., Geng, L.-S.: Review of deep learning based automatic segmentation for lung cancer radiotherapy. *Front. Oncol.* **11**, 717039 (2021)
6. Mansoor, A., Bagci, U., et al.: Segmentation and image analysis of abnormal lungs at CT: current approaches, challenges, and future trends. *Radiographics* **35**(4), 1056–1076 (2015)
7. Yoo, S.J., et al.: Automated lung segmentation on chest computed tomography images with extensive lung parenchymal abnormalities using a deep neural network. *Korean J. Radiol.* **22**(3), 476 (2021)
8. Hu, Q., et al.: An effective approach for CT lung segmentation using mask region based convolutional neural networks. *Artif. Intell. Med.* **103**, 101792 (2020)
9. Liu, C., Xie, W., Zhao, R., Pang, M.: Segmenting lung parenchyma from CT images with gray correlation based clustering. *IET Image Proc.* **17**(6), 1658–1667 (2023)
10. Han, J., Kamber, M.: *Data Mining: Concepts and Techniques*, 3rd edn. Elsevier, Burlington MA (2012)
11. OpenCV: K-Means Clustering in OpenCV. https://docs.opencv.org/3.4/d1/d5c/tutorial_py_kmeans_opencv.html (2023). Accessed 06 May 2023
12. Dougherty, G.: *Digital Image Processing for Medical Applications*. Cambridge University Press, Cambridge UK; New York (2009)
13. Tukey, J.: Nonlinear nonsuperposable methods for smoothing data. *Cong Rec EASCON* **74**, 673 (1974)
14. Matheron, G., Serra, J.: The birth of mathematical morphology. In: *International Symposium on Mathematical Morphology*, pp. 1–16 (2001)
15. Soille, P.: *Morphological Image Analysis: Principles and Applications*. Springer, Berlin Heidelberg (1999)
16. Dice, L.R.: Measures of the amount of ecologic association between species. *Ecology* **26**(3), 297–302 (1945)
17. Jaccard, P.: The distribution of the flora in the alpine zone. *New Phytol.* **11**(2), 37–50 (1912)
18. Kosub, S.: A note on the triangle inequality for the jaccard distance. *Pattern Recogn. Lett.* **120**, 36–38 (2016)
19. Gragera, A., Suppakitpaisarn, V.: Relaxed triangle inequality ratio of the sørensen dice and tversky indexes. *Theoret. Comput. Sci.* **718**, 37–45 (2018)
20. Depeursinge, A., Vargas, A., Platon, A., Geissbuhler, A., Poletti, P.A., Müller, H.: Building a reference multimedia database for interstitial lung diseases. *Comput. Med. Imaging Graph.. Med. Imaging Graph.* **36**(3), 227–238 (2012)