An automatic histogram-based initializing algorithm for K-means clustering in CT

Mengqiu Tian[1,2], Qiao Yang[1,2], Andreas Maier[2], Ingo Schasiepen[1], Nicole Maass[1], Matthias Elter[1]

[1]Siemens AG, H CP CV ME, Erlangen, Germany
[2]Pattern Recognition Lab, Friedrich-Alexander-University Erlangen-Nuremberg, Erlangen, Germany
tianmengqiu@gmail.com

Abstract. K-means clustering [1] has been widely used in various applications. One intrinsic limitation in K-means clustering is that the choice of initial clustering centroids may highly influence the performance of the algorithm. Some existing K-means initialization algorithms could generally achieve good results. However, in certain cases, such as CT images that contain several materials with similar gray-levels, such existing initialization algorithms will lead to poor performance in distinguishing those materials. We propose an automatic K-means initialization algorithm based on histogram analysis, which manages to overcome the aforementioned deficiency. Results demonstrate that our method achieves high efficiency in terms of finding starting points close to ground truth so that offers reliable segmentation results for CT images in aforementioned situation.

1 Introduction

Segmentation plays a critical role in many medical image processing applications, such as beam hardening correction in CT images. However, in practical cases, accurate object identification and separation are non-trivial tasks due to data acquisition and reconstruction artifacts.

Global thresholding has been widely used for object segmentation. However, most thresholding methods suffer from inaccurate detection of shapes and peaks in histogram due to noise and artifacts, thus leading to the process being difficult to fully automate. On the other hand, K-means clustering, which accepts classes with different shapes, has an intrinsic limitation that the computational efficiency highly depends on cluster initialization. Authors of [2] and [3] respectively compared 14 and 4 approaches for K-means clustering initialization. In [2] the scrambled midpoints and in [3] the Kaufman initialization algorithm have been reported to perform the best. However, in certain cases, the above two initialization methods for K-means clustering have the limitation to accurately distinguish materials with similar gray-levels in CT images. For example, in application such as segmenting hearing aid containing air, plastic and metal components, those methods are not sensitive enough to separate air and plastic

which have much closer densities compared to metal. We propose a novel K-means initialization algorithm, which offers better sensitivity in distinguishing similar materials, and meanwhile yields better performance for general datasets . The new algorithm achieves automatic initialization of cluster centroids for K-means with prior-knowledge from histogram. In Section 3, we compare the proposed algorithm with the scrambled midpoints and the Kaufman initialization algorithms with respect to initialization accuracy, and Otsu's method, a classical histogram-based thresholding method, to compare segmentation performance [4,5].

2 Materials and Methods

The algorithm's main idea is to detect the peaks in histogram and to employ the corresponding gray-levels as initial cluster centroids for K-means clustering. For the first centroid, the gray-level of the highest peak in the histogram is selected. Then the local maximum with the largest weighted distance to all other known centroids is chosen as a new centroid.

2.1 Algorithm

For the distance metric, the product of the gray-level difference $d_{n,i}$ between centroid C_n and local maximum i, and the height h_i of the local maximum is used. This process is iteratively repeated until the desired number of centroids is found. We use the product (instead of the sum) of all weighted distances as criterion, since iff. all $(h_i \cdot d_{n,i})$ were large, the product would be maximized. The flowchart of our algorithm is depicted in Fig. 1.

1. Calculate the image histogram.
2. Detect all local maxima in the histogram. We assume a total number of I local maxima are detected.
3. Choose the global maximum of the histogram as the first centroid C_1.
4. Identify the remaining centroids as follows:
   (a) Calculate
   $$P_{N+1}(i) = \prod_{n=1}^{N} (h_i \cdot d_{n,i}) \tag{1}$$
   for each local maximum $(i = 1, 2, ..., I)$. N is the number of already found centroids. The computational complexity can be reduced by using a recursive implementation:
   $$P_2(i) = h_i \cdot d_{1,i} \tag{2}$$
   $$P_{N+1}(i) = P_N(i) \cdot (h_i \cdot d_{N,i}) \tag{3}$$
   which employs similar idea of Viterbi algorithm [6].
   (b) Then the new centroid $C_{N+1} = \arg\max_i(P_{N+1}(i)), N = N + 1$.
5. Repeat 4 until N equals the user-specified number of classes K.

2.2 Experiments

The proposed method was evaluated with ten datasets containing simulated data and CT scans. Three simulated datasets were obtained using DRASIM (Siemens Healthcare, Forchheim, Germany), and seven real CT datasets were obtained from industrial and medical CT scanners (industrial datasets: Siemens Healthcare, Munich, Germany; medical datasets: Siemens Healthcare, Erlangen, Germany; Dataset I: jaw phantom; II: head phantom; III: aluminum and iron cylinders phantom IV: hearing-aid; V: computer mouse; VI: aluminum component; VII: CD spindle; VIII: elbow; IX: knee; X: head) All datasets were reconstructed using filtered back projection (FBP) on a $512 \times 512 \times 512$ grid, except for dataset IX which uses FBP on a $384 \times 384 \times 384$ grid. We segmented all these ten datasets, and read out the segmentation results.

We use the *Normalized Sum of Absolute Difference*(NSAD) to measure the proximity of initial cluster centroids to ground truth. The smaller the NSAD value is, the closer the initial cluster centroids are to the ground truth. If NSAD is too large, it can happen that K-means clustering algorithm does not converge to ground truth.

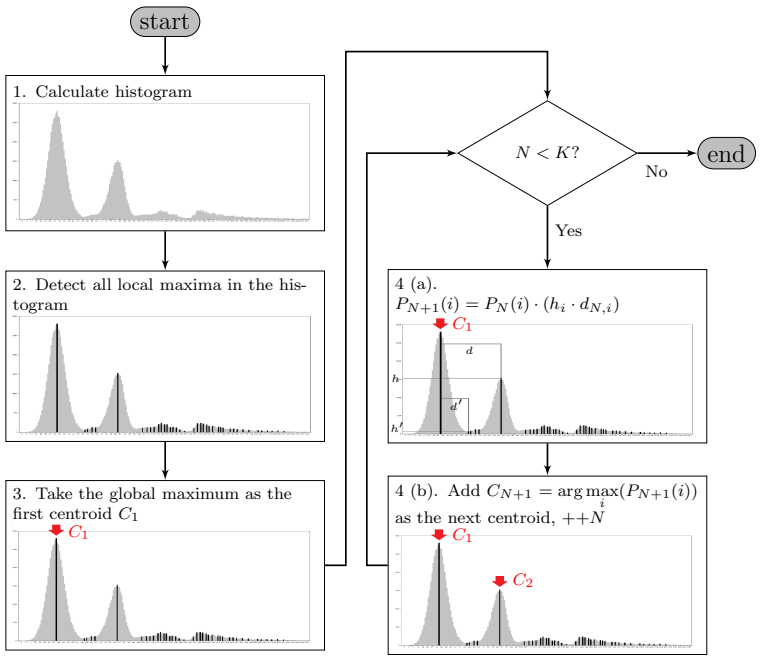


Fig. 1. Flowchart of the proposed algorithm.
(X-axis of histogram: bin index; Y-axis of histogram: number of pixels)

We first calculated the normalized distance between the initial cluster centroids given by initialization algorithms and ground truth:

$$d[k] = \frac{|\text{Ground truth}[k] - \text{Initial cluster centroid}[k]|}{\Delta t} \tag{4}$$

Δt is the width of each histogram bin: $\Delta t = (v_{max} - v_{min})/L$, where v_{min} and v_{max} are the minimal and maximal gray-levels of the image respectively, and L is the number of histogram bins. 512 bins are applied in our experiments for all datasets. Then we have

$$NSAD = \sum_{k=1}^{K} d[k] \tag{5}$$

3 Results

In this section, we present the results from a series of experiments, comparing our algorithm with alternative approaches in terms of segmentation accuracy and the proximity of starting points to ground truth.

3.1 Analysis of segmentation result

The *percentage of misclassified pixels* (pMP) are used to measure the performance of three initialization algorithms and multi-level Otsu's method. Fig. 2 shows that our proposed method results in the best clustering results with minimal pMP in our datasets.

3.2 Analysis of proximity of starting points to ground truth

Tab. 1 shows a summary of NSAD for ten datasets ranging from simulated data to experimental CT data. For experimental datasets, human observers selected

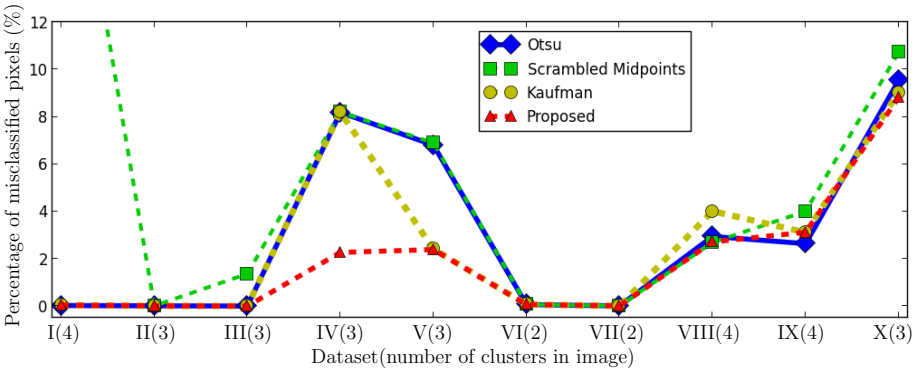


Fig. 2. Number of misclassified pixels in each dataset (%).

Table 1. Normalized Sum of Absolute Difference (NSAD).

Dataset	I	II	III	IV	V	VI	VII	VIII	IX	X
Scrambled Midpoints	525	122	331	633	690	90.8	135	514	484	307
Kaufman	30.9	36.3	55.5	331	77.6	1.67	30.2	41.5	50.9	46.8
Proposed	4.48	6.33×10^{-10}	3.19×10^{-10}	17.0	1.0	10.0	2.0	1.0	212	105

a given number of peaks from histograms, and used the average value of corresponding gray values as ground truth. The highlighted terms show the smallest NSAD value for each dataset. It can be seen that, for most datasets the proposed algorithm outperforms others.

We depict the segmentation results from initial cluster centroids of three datasets (one simulated dataset I and two real datasets IV and VIII) in Fig. 3, in order to compare the performance of selected initial centroids with ground truth. It can be seen that the segmentation results of proposed algorithm are already very close to the segmentation results based on ground truth, while scrambled midpoints and Kaufman method could not perform as well as expected. The reason is when datasets consisting of similarly dense objects, the latter two methods have lower sensitivity to detect histogram peaks. For example, in hearing-aid dataset, air and plastic have close density values, resulting in the peaks of air

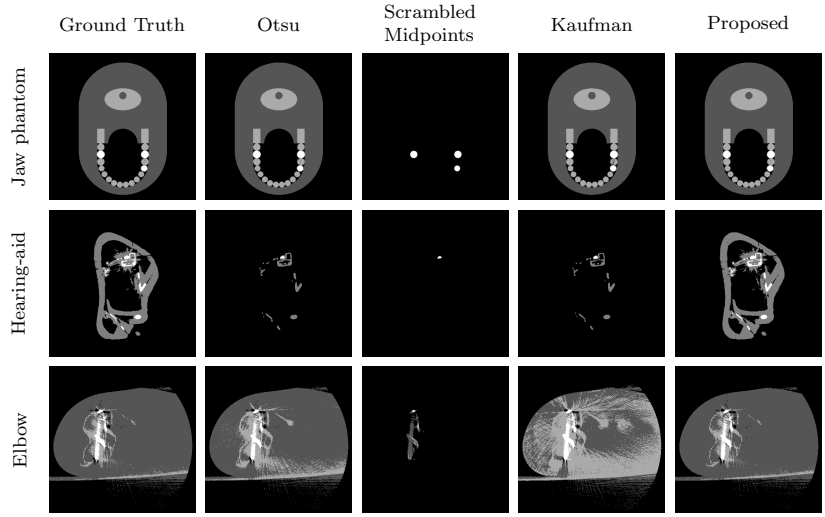
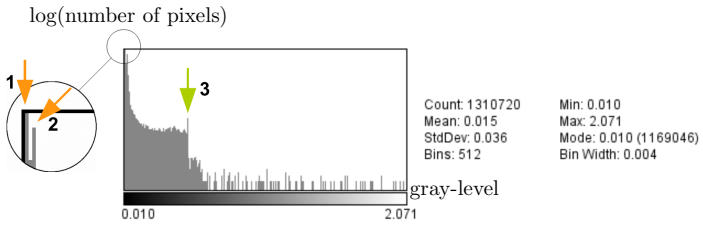


Fig. 3. Segmentation results analysis: Column 1: Segmentation results based on ground truth; Column 2: Segmentation results of multi-level Otsu Method; Column 3-5: Segmentation results of K-means initial clustering centroids for scrambled midpoints, Kaufman and proposed method.

Fig. 4. Logarithm histogram of hearing-aid.



and plastic (orange arrows 1 and 2) located very close to each other compared to the peak of metal (green arrow 3) (Fig. 4). Moreover, our algorithm leads to better segmentation results than multi-level Otsu's algorithm, while multi-level Otsu's algorithm is very time consuming. The computational complexity of the proposed algorithm is $\mathcal{O}(KL)$, while it is $\mathcal{O}(L^{K-1})$ for multi-level Otsu's method [5].

4 Discussion

In this paper, we compared the proposed method with several existing initialization methods on a variety of datasets. Experimental results demonstrate that the proposed algorithm achieves better accuracy and closer proximity of initial centroids to ground truth. However, even if K-means algorithm is properly initialized, it might be insufficient for medical purpose since in medical CT images, different organs at different location could yield same gray levels. In such cases, the segmentation could not based only on gray levels, but also should take features or spatial information into account.

References

1. MacQueen J. Some methods for classification and analysis of multivariate observations. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. 1967;1:281–297.
2. Robinson F, Apon A, Brewer D, Dowdy L, Hoffman D, Lu B. Initial Starting Point Analysis for K-Means Clustering: A Case Study. Proceedings of ALAR 2006 Conference on Applied Research in Information Technology. 2006 March;.
3. Pena JM, Lozano JA, Larranaga P. An empirical comparison of four initialization methods for the K-Means algorithm. Pattern Recognit Lett. 1999;20:1027–1040.
4. Otsu N. A Threshold Selection Method from Gray-Level Histograms. IEEE Trans Syst Man Cybern. 1979 January;SMC-9:62–66.
5. Liao PS, Chen TS, Chung PC. A fast algorithm for Multilevel Thresholding. Journal of information science and engineering. 2001;17:713–727.
6. Huang X, Acero A, Hon HW. Spoken Language Processing: A Guide to Theory, Algorithm, and System Development. 1st ed. Upper Saddle River, NJ, USA: Prentice Hall PTR; 2001.