ELECTRONICS AND ELECTRICAL ENGINEERING
ISSN 1392 – 1215 2011. No. 4(110)
ELEKTRONIKA IR ELEKTROTECHNIKA

SYSTEM ENGINEERING, COMPUTER TECHNOLOGY
T 120
SISTEMŲ INŽINERIJA, KOMPIUTERINĖS TECHNOLOGIJOS

Parallel Algorithm Evaluation in the Image and Clustering Processing

C. Pughineanu, I. Balan
Faculty of Electrical Engineering and Computer Science, Ştefan cel Mare University of Suceava,
Str. Universitatii nr.13, 720229 Suceava, România, phone: +40 230 524801, e-mail: secretariat@eed.usv.ro

Introduction

The ISODATA algorithm parallelized [1–16] with the two RGB and RGBXY models will be used to reduce the number of the colors, the color images obtained as such will be converted into grayscale images.

The segmentation of the grayscale images [2] will be achieved with the parallelized k-means algorithm [3], a model advanced by (1967) and the modified parallel ISODATA, which mainly is similar to the k-means one, the difference being the fact that here the number of clusters to be determined may be automatically modified in the time of iteration by similar group fusion and division of the groups with great standard deviations.

The results of the experiments were obtained by the execution of the algorithms on a cluster that has the architecture presented in Fig. 1. This cluster, of the „Ştefan cel Mare" University of Suceava, is formed of 28 nodes, each node being represented by a HS21 blade server with 2 Xeon Quad Core E5345 2.33GHz/1333MHz/8MB L2 processors. These nodes are linked among them by a communication network of 1Gbps to perform the calculations by a 1Gpbs network for management, too [12–16].

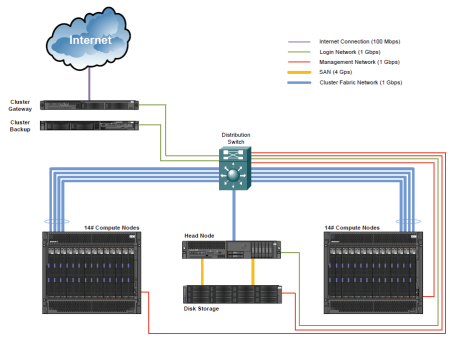


Fig. 1. The cluster of the „Ştefan cel Mare" University of Suceava

The k-means algorithm

The k-means algorithm aims to identify distinct k groups (clusters), so as the data of each class to be similar enough [4]-[5]. Every class will have a representative considered to be the centre of the class. These centers are random defined. The affiliation of each point to a cluster is determined. Every segmentation by clustering [6] will be followed by the labeling of the segmented image to identify the individual objects. If the points do not change the clusters they belong to, then the grouping is finished. In the opposite situation, the centers for each cluster must be recalculated due to the points that are part of it. After the calculation of these centers, the affiliation of each point to a cluster is determined. As such, a curl is generated. At every iteration in the curl, the clusters change their centre. The issue from the curl may be achieved in the moment the clusters do not change their centers.

If the data set [11–12] contains n points $(x_1, x_2, ...x_n)$, in which every point is considered as an observation vector of the dimension d $(d=3)$, then the k-means clustering involves dividing of this set into k partitions (classes) disjunct and non-zero $(k<n)$ $S=\{S_1,S_2,...S_k\}$ so that the similarity function minimizes [10] (sum of the square distances from the cluster inside) [5, 6]

$$E = \arg\min_{s} \sum_{i=1}^{k} \sum_{x_j \in S_i} \left\| x_j - c_i \right\|^2 , \qquad (1)$$

where c_i is the mean of S_i .

The ISODATA algorithm

An important role in the segmentation process is constituted by the initial choice of the number of classes. A wrong choice causes the errors of over-segmentation or sub-segmentation. The over-segmentation may be corrected by the further unification of the objects already determined and, thus, it is preferable to choose a greater k than necessary.

Initially one wants the segmentation of the image into k clusters, but he ISODATA algorithm [7] allows the number of the clusters to be automatically modified during an iteration by fusing the similar groups and dividing those with great standard deviations. At the beginning, the k centers are arbitrarily chosen, then each point is attributed to the closest cluster. If there are clusters with fewer Θ_N points (threshold value which a minimum number of

samples from each cluster may have) then these are far off, each centre remaining to be recalculated [8]. If there are too few clusters, the first step is taken to separate, by determination, the vector of the standard deviations for each cluster. The centre of the cluster is divided into two new cluster centers, then $k \leftarrow k+1$ is also erased. If there are too many clusters, the first step to fusion is taken. The fusing centers and $k \leftarrow k-1$ are erased. The algorithm is finished if the maximum iteration number I is attained [9]. Thus, the algorithm is restarted/retaken by attributing each point to the closest cluster.

The ISODATA algorithm, used to reduce the colors, receives, at entering, an image in the BMP format, using as a model the RGB color system and optionally the position of every RGBXY pixel. The pixels, which will present in the image a similar color and position, will be regrouped and each pixel will be identified by the number that shows the color class to which it is the most similar. The program contains two functions *isodataRGBXY* and *isodataRGB,* the first modifies the image using the ISODATA algorithm, considering both the color and the pixel position, while as the second uses for grouping only the color of each pixel.

Parallel implementation of the k-means and ISODATA algorithms

In the parallel implementation will be partitioned the calculations that will be achieved by associating every operation with the data it operates with. This partitioning produces a number of tasks, each containing data and a set of operations for them. There might be cases in which an operation must be performed on the data of the different tasks. In these cases, it is necessary to achieve the communication among these tasks [8, 10, 11].

For the present case we will use the strategy of achieving the same set of calculations on different data sets. Thus, the data are divided into equal parts, their number being equal to the number of processes to be executed in parallel. The implementation of these algorithms is achieved with the help of the MPL standard.

Experimental results

The ISODATA is tested on color images to see which of the two ways, RGBXY or RGB, is more efficient. We may notice that RGBXY way offers better results almost all the time, although there also are some "extreme" cases (when the image is reduced to a small number of colors and the original image has a small number of colors, too), when the RGB modality succeeds to capture a greater number of details and thus is more efficient.

Image 7 (Fig. 8) is one of those "extreme" cases. One may notice that the RGB processing way is by far better than the RGBXY one.

As a conclusion, for the color images we shall use the RGBXY variant and the white-black ones, variant RGB.

Image 4 (Fig. 5) original and image 5 (Fig. 6) processed in the RGBXY way will be transformed into grayscale image then segmented and the interest zone will be extracted, with parallel k-means and ISODATA algorithms.

Fig. 2. Image 1 original

Fig. 3. Image 2 processed in the RGBXY modality

Fig. 4. Image 3 processed in the RGB modality reducing to 20 colors with 10 iterations

Fig. 5. Image 4 original

Fig. 6. Image 5 processed in the RGBXY modality

Fig. 7. Image 6 processed in the RGB modality reducing to 24 colors with 30 iterations

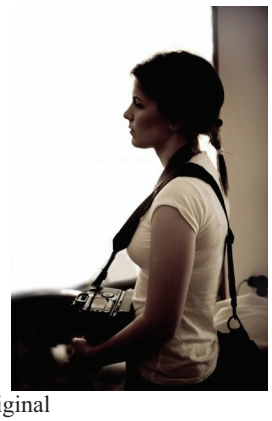
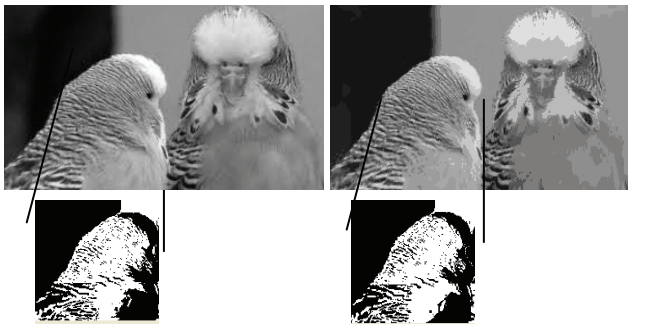


Fig. 8. Image 7 original

Fig. 9. Image 8 processed in the RGB modality

Fig. 10. Image 9 processed in the RGBXY modality reducing to 2 colors with 10 iterations

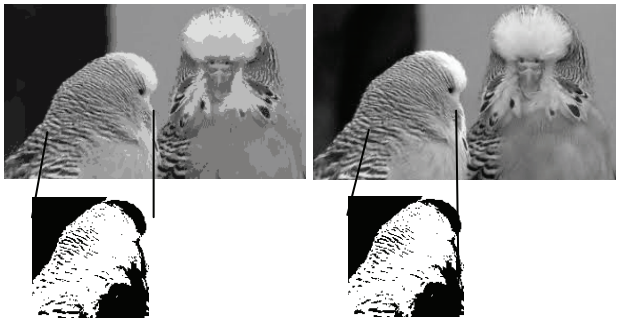


Interest zone extraction (class 1) from image 4, using the k-means algorithm

Interest zone extraction (class 1) from image 5, using the k-means algorithm

Fig. 11. Interest zone extraction (class 1)

The images were segmented into 5 classes with k-means algorithms and the image belonging to class 1 was extracted.



Interest zone extraction (class 0) in image 4, using the ISODATA algorithm

Interest zone extraction (class 0) in image 5, using the ISODATA algorithm

Fig. 12. Interest zone extraction (class 0)

Applying the parallelized ISODATA algorithm to the interest set, namely the image in figure 4, segments the pixels also into 5 classes, extracted being the image belonging to class 0 (Fig. 12).

Comparing the extracted interest zones, with the two parallelized algorithms k-means and ISODATA using image 4 original (Fig. 5), we notice that the pixel group with parallelized ISODATA algorithm offers much better results.

Conclusions

In this article, we presented the parallel algorithm to segment the k-means images and the parallel ISODATA algorithm with the two variants both for image segmentation and for color number reduction.

The increase of the volume of information of the image type asks for stocking, which implies space. That is why we proved that reducing the number of colors does not influence the quality of the results as segmentation follow-up.

We proved that the parallel ISODATA algorithm used in image segmentation is much better compared to the k-means algorithm, the zone/area extracted with ISODATA is of greater similarity to the one from the original image.

Acknowledgements

I would like to thank Faculty of Electrical Engineering and Computer Science, Ştefan cel Mare University of Suceava, Romania for allowing me to use the computer facilities in my experiment.

References

1. **Feng Z., Zhou B., Shen J.**, A parallel hierarchical clustering algorithm for PCs cluster system // Neurocomputing, 2007. – Vol. 70. – Iss. 4–6. – P.809–818.

2. **Berry W. M., Castellanos M.** Survey of Text Mining II: Clustering, Classification, and Retrieval. – Springer, 2008.
3. **Kanungo T., Mount D. M., Netanyahu N. S., Piatko C. D., Silverman R., Wu A. Y.** An efficient k–means clustering algorithm: analysis and implementation // IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002. – Vol. 24. – Iss. 7. – P. 881–892.
4. **Teuvo K.** Self–organization and associative memory. – Springer–Verlag Berlin Heidelberg New York London Paris Tokyo, 1988.
5. **Pentiuc S. G., Schipor O. A., Danubianu M., Schipor M.D., Tobolcea I**. Speech Therapy Programs for a Computer Aided Therapy System // Electronics and Electrical Engineering. – Kaunas: Technologija, 2010. – No. 7(103). – P. 87–90.
6. **Pentiuc S. G.** Aplicații ale recunoașterii formelor în diagnosticul automat. – Editura Tehnică, Bucureşti, 1997.
7. **Memarsadeghi N., Mount M. D., Netanyahu S. N., Le Moigne J.** A fast implementation of the ISODATA clustering algorithm // International Journal of Computational Geometry and Applications (IJCGA), 2007. – No. 17(1). – P. 71–103.
8. **Zamir O., Etzioni O.** Web Document Clustering: A Feasibility Demonstration // Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1998. – P. 46–54.
9. **Cover T. M., Hart P. E.** Nearest Neighbor Pattern Classification // IEEE Trans. on Information Theory, 1967. – IT–13, 1. – P. 21–26,
10. **Rata G., Rata M., Filote C., Strugaru C**. Theoretical and Experimental Aspects Concerning Fourier and Wavelet Analysis for Deforming Consumers in Power Network // Electronics and Electrical Engineering. – Kaunas: Technologija, 2010. – No. 1(97). – P. 62–66.
11. **Dambrauskas A., Rinkevicius V**. Algorithmic Methods of Variational Calculus // Electronics and Electrical Engineering. – Kaunas: Technologija, 2008. – No. 5(85). – P. 25–28.
12. **Ungurean I.** Job Scheduling Algorithm based on Dynamic Management of Resources Provided by Grid Computing Systems // Electronics and Electrical Engineering. – Kaunas: Technologija, 2010. – No. 7(103). – P. 57–60.
13. **Vahdat–Nejad H., Zamanifar K.**, A New Randomized Algorithm for Handling Scheduling Conflicts in Grids // Advances in Electrical and Computer Engineering, 2009. – Vol. 9, No. 3. – P. 22–26.
14. **Ciufudean C., Filote C.** Safety Discrete Event Models for Holonic Cyclic Manufacturing Systems // The 4th International Conference on Industrial Applications of Holonic and Multi–Agent Systems. – HoloMAS, Linz, Springer Verlag, 2009. – P. 225 – 233.
15. **Ciufudean C., Filote C., Buzduga C.** Electronic Device for Monitoring Electrical and Non–electrical Measurands // Electronics and Electrical Engineering. – Kaunas: Technologija, 2009. – No. 7(95). – P. 51–54.
16. **Ciufudean C., Larionescu A., Filote C.** Equivalent Structure for Nonlinear–Signal Cable Networks // Electronics and Electrical Engineering. – Kaunas: Technologija, 2009. – No. 5(93). – P. 65–68.

Received 2011 02 11

C. Pughineanu, I. Balan. Parallel Algorithm Evaluation in the Image and Clustering Processing // Electronics and Electrical Engineering. – Kaunas: Technologija, 2011. – No. 4(110). – P. 89–92.

The increase of the information volume of the image type in the greatest part of the domains asks for the introduction of some storage and efficient recovery methods of the available data due to content. Unfortunately, the progress registered in the field of the multimedia databases with digital images is not remarkable as being outdated by info explosion. The article proposes compression algorithms aiming to reduce the quantity of data necessary to represent an image and the necessary clustering algorithms, namely the k-means algorithm and ISODATA, which were parallelized both from the point of view of the extracted areas and the execution time. The experimental results were obtained by the implementation of the algorithms using the MPI standard and their execution on a cluster. Ill. 12, bibl. 16 (in English; abstracts in English and Lithuanian).

C. Pughineanu, I. Balan. Lygiagrečiojo algoritmo įvertinimas vaizdų apdorojimo ir informacijos grupavimo sistemose // Elektronika ir elektrotechnika. – Kaunas: Technologija, 2011. – Nr. 4(110). – P. 89–92.

Didesniam vaizdinės informacijos kiekiui išsaugoti reikia didesnio failo. Reikalingi nauji būdai ir metodai tokiai informacijai efektyviai išsaugoti ir atkurti. Straipsnyje siūlomas glaudinimo algoritmas, kuris pasižymi duomenų kiekio mažinimu, reikalingu vaizdinei informacijai atkurti, ir tiems duomenims grupuoti. Analizei panaudoti vadinamieji k ir ISODATA algoritmai. Pateikti eksperimentiniai rezultatatai. Il. 12, bibl. 16 (anglų kalba; santraukos anglų ir lietuvių k.).