NOVEL HIERARCHICAL K-MEANS CLUSTERING BASED LOSSLESS DATA COMPRESSION SCHEME

Dr. J. Satheesh Kumar
Department of ECE
Hindusthan College of Engineering and Technology
Coimbatore, Tamilnadu, India

S.V. Krithika,
Department of EEE,
V.S.B College of Engineering,
Coimbatore, Tamilnadu, India

Abstract— **Hierarchical K-means is an important clustering method that is widely applied in different applications such as wireless networks, data mining and image processing. Now this method can be applied in very large scale integration (VLSI) domain. It is the process of creating an integrated circuit. This new approach can be determined initial from the centroid for k-means. The hierarchical clustering is most preferable because of flexibility. In this paper a Hierarchical K-means clustering (NHK) based golomb coding technique is implemented on Field Programmable Gate Array (FPGA) for reconfigurable system design. Golomb coding is the lossless data compression method using a family of data compression codes. Lossless data compression is used in many applications. For example, it is used in the ZIP file format. Golomb coding is a well-known technique for a lossless data compression. The coding plan improvement in FPGA uses the Verilog HDL. Progressive memory is utilized to store an aggregate of 2046 centroid vectors and offer a limit of 10 centroid vectors to the area all the while.**

Keywords — Clustering method, Hierarchical K-means architecture, FPGA, Golomb coding.

I. INTRODUCTION

Field Programmable Gate Arrays (FPGAs) are widely Field-programmable gate arrays (FPGAs) are widely used in reconfigurable silicon devices for complex and safety critical Applications. Seeing as, the configuration information for FPGA devices must be put away by inner memory as bit streams. The memory size and access rapid across the board development of computerized advances, for example, Internet access, video player, computerized TV, video, games, interactive media and cell phones have expanded the interest for high stockpiling and furthermore bandwidth so as to fit the developing needs S. W. Golomb et al. (1966). Hierarchical clustering can joins nearby points into a group and afterward progressively adds close by focuses to the closest gathering. This has called upon for the need of proficient information pressure procedures, where the first information is compacted. This procedure diminishes the transmission capacity required for information move. The target of this task work is to structure a Novel Hierarchical

K-Means Clustering (NHK) for FPGA memory configuration to decrease the calculation time and equipment cost. Golomb coding I. H. Witten and et al. (1987) is equipped for packing bigger measured information into littler estimated information while as yet permitting the first information to be reproduced back after decompression. In addition, there is other elite lossless pressure calculation, for example, number-crunching coding A. Chandra and et al. (2001) and so on. In any case, this calculation includes higher plan multifaceted nature and computational burden. Golomb coding applications remembers the utilization for System on-chip (SoC) test-information pressure framework as introduced in T.W. Chen and et al. (2008).In the conventional K-Means, the calculation multifaceted nature is corresponding to the bunch number. By the by, the current K-Means models can't fulfil the expenses of both the computational time and equipment territory. At the point when the group number is fixed, the computational time and the territory costs are in reverse extent. As such, if the computational time is decreased, the number of equal preparing components will increment. The expenses in installed frameworks are consistently significant, and the battle between the computational time and the equipment zone becomes serious particularly as the bunch number increments. To defeat this issue, another calculation dependent on Hierarchical K-Means (HK-Means) is proposed. This new design incorporates a various levelled memory structure to store the cluster centroids. The commitments of this work are twofold. The one is the engineering investigation of HK-Means calculation, and the other is configuration space investigation to store the high group number, Among our works Tung-Shouchen and et al. (2005) the proposed engineering has the lowest entryway tally per bunch to help the most elevated group number, and it is appropriate for installed frameworks which mainly centre around sight and sound substance examination in the following generation. This paper is sorted out as follows. The examination foundation is talked about in area II. The calculation of the HK-Cluster is depicted in segment III, and the segment IV present the subtleties of golomb coding, and the Section V presents the reproduction consequences of the calculations produced so as to demonstrate their legitimacy. Ultimately segment VI closes the paper.

II. RESEARCH BACKGROUND

Among the information investigation grouping techniques, K-Means and Hierarchical clustering are specialists' great apparatuses today. Be that as it may, every one of these conventional bunching techniques has its restrictions. In this investigation, we present another strategy, various leveled K-Means controlling troublesome or agglomerative methodology. The various leveled K-Means right off the bat utilizes K-means' calculation in each group to decide K cluster while working and afterward utilizes it on progressive grouping method to abbreviate consolidating bunches time while creating a tree-like. The outcome demonstrates disruptive various leveled K-clustering is better than progressive grouping on bunch quality and is better than K-Means Clustering on computational speed David casasemt and et al. (2006). A double various leveled order structure to address the multi-class arrangement issue with another progressive plan technique, k-Clustering SVRM (support vector representation machine) grouping. This enormously enhances our earlier IJCNN various leveled structure. At every hub in the progressive system, we apply the SVRDM (support vector representation and discrimination machine) classifier, which offers speculation and great dismissal capacity Yi pan,R.Harrison et al. (2006) K-Means is a grouping calculation that is generally applied in numerous fields, including design order, sight and sound examination, and picture recovery. Because of constant prerequisites of picture division in inserted frameworks, it is important to quicken K-Means calculation by equipment executions. The commitment of this paper incorporates a progression of K-Means equipment investigations and a recently proposed SIP for picture division in SoC situations. K-cluster bunching requires a predefined number of groups ahead of time and picks introductory centroids arbitrarily also; it is delicate to anomalies. K-cluster grouping calculation, the calculation is hard to be applied across the board prevalently attributable to its high computational multifaceted nature. So a novel grouping calculation dependent on various leveled and K-cluster bunching, this has great computational multifaceted nature. Data pressure calculations have been broadly utilized in numerous territories to fulfill the need of capacity and move of huge size information. Shenfengchen and et al.(2010). The majorities of the information pressure calculations view the contribution as a grouping of double numbers and speak to the compacted information additionally as a parallel succession. Golomb coding is utilized to improve the exhibitions of the Golomb group of codes, including Golomb coding (GC), Golomb–Rice coding (GRC), exp-Golomb coding (EGC), and half and Hybrid Golomb coding (HGC), for information pressure. The Golomb group of codes is unmistakably fit to the handling of information with mathematical dissemination. Lossless information pressure is the way toward encoding (compacting) a group of information into a littler collection of information which can sometime in the not too distant future be particularly decoded (decompressed) back to the first information, while the decompressed information are

some estimate of the first information in lossy pressure. Lossless Compression is a key in text or parallel document pressure for putting away and transmission. The pressure methods which are lossless in nature and can decrease the information characters in a considerable manner. Both the calculations are versatile and there is no need of moving any additional data from encoder to decoder.
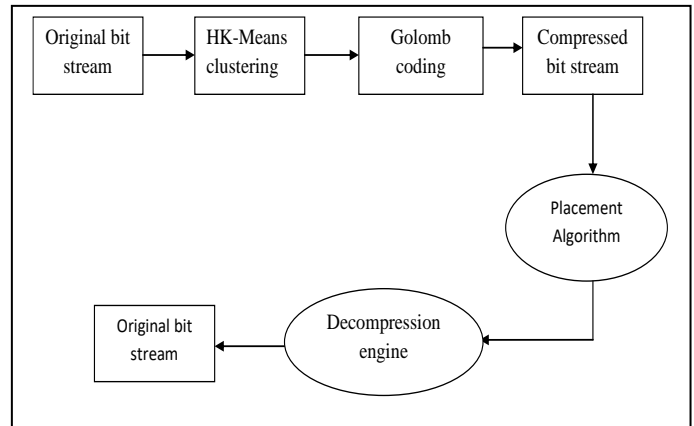
BLOCK DIAGRAM:

Figure 1: HK-means clustering decode- aware bit stream compression with golomb coding.

Figure 1: shows the Block diagram of Proposed HK-Means clustering Decode-aware bit stream compression with Golomb coding. Hierarchical k-means structure gathers the group centroids, and twofold tree crossings are utilized to register the closest centroid activities to frame the bunches. Compacted bit stream is communicated from the memory to the position calculation and afterward to the decompression Engine. Thus the first arrangement bitstream is created by decompression.

III. HK-MEANS CLUSTERING ALGORITHM

Clustering algorithm are utilized to part an informational index into a few gatherings (i.e., clusters), with the goal that the article in a similar gathering are as comparative as could reasonably be expected. The most main stream Clustering algorithm is K-means clustering and Hierarchical Clustering. K-means clustering is a dividing strategy utilized for parting a dataset into a lot of K clusters. Various levelled grouping is an elective way to deal with K-means bunching for distinguishing bunching in the dataset by utilizing pair wise separation lattice between perceptions as grouping standards. Hierarchical clustering grouping creates a tree structure in the bunching procedure, and there are top-down and base up grouping draws near. Since there are no bunching strategies that are reasonable for all the issues, numerous correlative grouping calculations that consolidate K-Means with Hierarchical clustering are proposed. In the k-means clustering technique on the off chance that we send the 4 piece information the comparing

bit can be group around then, we can't group the another size of information, to evade this difficult we have move to HK-cluster grouping strategy. In this technique all size of information can bunch at the same time zhouyong et al. (2006). This is the primary bit of leeway behind K-means group. Our strategy is planned a Novel Hierarchical K-means clustering procedure for productive piece stream pressure to improve both pressure and decompression efficiencies. In this technique pick some K to the expanding factor. This characterizes the quantity of group at each degree of the bunching chain of importance, and afterward we group the arrangement of focuses into K-cluster utilizing a standard K-implies calculation. N. Chehata and et al. (2008). In this paper, HK-means alludes to a top-down and disruptive Hierarchical K-means clustering algorithm that embraces K-means clustering with group number in each stage. The idea of HK-means is typically applied to lessen the computational time in the software algorithm when the group is huge. In the proposed HK-implies grouping, this is the strategy that is reasonable for equipment plan. The group are part into two recursively in the Euclidean space; the K-means bunching with it is performed locally in each level dependent on the grouping consequence of past level. The progression in the HK-means clustering is expressed as follows.

Figure 2: The step in the HK-means clustering is stated as follows as:

Step 1: At first; perform K-Means on all the input vectors with bunch number, and the information vectors are isolated into two kid groups with succession number 0 and 1. Since it is the initial step, at any rate two groups will be created in the HK-Means clustering process. Set the current level n-1 and the current sequence number t-0 . Go to step2.
Step 2: The stopping sequence number is algorithm by using the following equation:

$$T_S = K - 2^{n-1} - 1$$

Where 'n' denotes the current level. The process of HK-Means clustering ends when the following two conditions are satisfied:

$$t = T_S$$

and

$$n = [log_2 K]$$

Where 't' represents the current arrangement number. At the point when the clustering procedure ends, set the last level $n_{final} = n$ and the last halting succession number Ts(final)= Ts for the grouping results. In any case go to stage 3.

Step 3: If the current arrangement number t in the (n-1)th level is equivalent to $2_{(n-1)}$ 1, it implies that 2_n bunches are created in the current level. Go to the following level by expanding n by 1, set t=0 and go to Step 4. Something else, increment t by 1 without changing the estimation of n and go to Step 4.

Stage 4: Presently performs K-Means on the input vectors that have a place with the T [th] cluster in the (n-1) [th] level. The group number of K-Means is set to k=2. After the K-Means After the K-Means clustering dependent on the Tth group in the (n-1)[th] level is done, two new bunches with arrangement number 2t and 2t+1 are produced in the current level (the nth level). Go to Step 2.

IV. GOLOMB CODING

The advancement of Golomb coding calculations for information pressure and decompression and their execution in Field Programmable Gate Array (FPGA) is introduced. So as to have effortlessness being developed and testing, the golomb coding boundary m is set. The m characterizes the code structure. In this way, picking the m boundary chooses variable length code structure which will have direct effect on the pressure effectiveness. Encoder is the way toward changing over various contributions of information into single yield of information for a portion of the reasons, for example, normalization, speed, pressure, and so forth. In our task the encoder can pack the mass information into a straightforward yield. After the way toward encoding the scrambled information can goes under the way toward unraveling. Decoder is the converse procedure of encoder i.e., changing over the scrambled content into the first plain content. The way toward disentangling can be utilized to recuperate the first information from boisterous information; by this procedure we can get a unique piece stream.
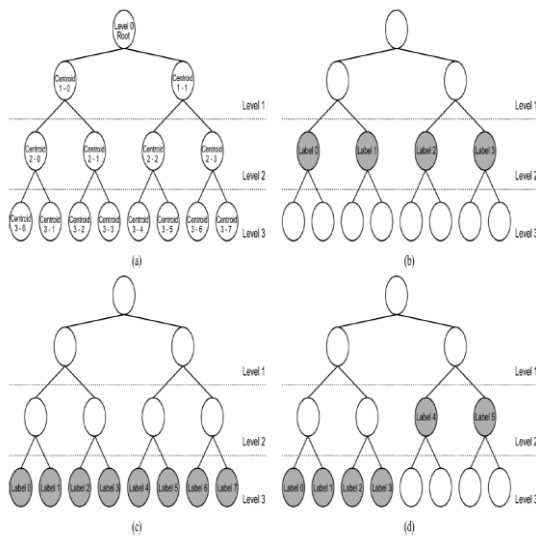
Group	Run-length	Group prefix	Tail	Codeword
A1	0	0	00	000
	1		01	001
	2		10	010
	3		11	011
A2	4	10	00	1000
	5		01	1001
	6		10	1010
	7		11	1011
A3	8	110	00	11000
	9		01	11001
	10		10	11010
	11		11	11011

Figure 3: Golomb coding example with parameter m=4.

Through reproduction of golomb coding, it tends to be reasoned that the calculation of golomb coding on the FPGA had been effectively evolved. This had been demonstrated by contrasting the outcomes created utilizing recreation of the golomb encoder with the normal outcomes which indicated both of the outcomes were indistinguishable. The utilization of encoded information as the info information for decoder figured out how to produce back the first information demonstrated that the golomb decoder framework was likewise effectively evolved. Thus, the entire arrangement of golomb coding was finished. It is a notable method because of its lower multifaceted nature.

V. RESULTS AND DISCUSSION:

Power summary:	I(mA)	P(mW)
Total estimated power consumption:		56
Vccint 1.20V:	15	19
Vccaux 2.50V:	15	38
Vcco25 2.50V:	0	0
Clocks:	0	0
Inputs:	0	0
Logic:	0	0
Outputs:		
Vcco25	0	0
Signals:	0	0
Quiescent Vccint 1.20V:	15	19
Quiescent Vccaux 2.50V:	15	38

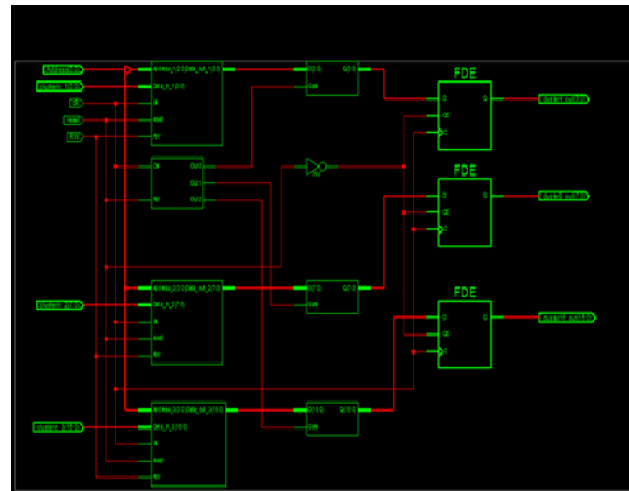Figure 4: power analysis of HK-means clustering

Figure 5: RTL schematic diagram of clustering process.

VI. CONCLUSION

The current k means algorithm may break down the information or straightforwardly bunch the information. In this paper we proposed a HK-means clustering method which is utilized to group the information in the wake of breaking down the size of the info bitstream. Through reproduction of HK-means cluster it very well may be presumed that the calculation of HK-means clustering coding on FPGA had been effectively evolved. Our structure objectives in proposing such a calculation centers around limiting the bitstream size, the memory necessity of FPGA setup, and separating the computational multifaceted nature. Proposed innovation utilized for information pressure to lessen memory consumed space in FPGA.

VII. REFERENCE

[1] S. W. Golomb (1966) "Run Length Encodings," IEEE Transactions on Information Theory, vol. 12, (pp.399-401).

[2] I. H. Witten, R. M. Neal, and J. G. Cleary, (1987) "Arithmetic Coding for Data Compression" CACMJournal,vol. 30, no. 6 (pp.520–540).

[3] A. Chandra and K. Chakrabarty, (2001) "System on-a-chip test data compression and decompression architectures based on Golomb codes," IEEE Trans.Computer-Aided Design, vol. 20 (pp.355-368).

[4] T.W. Chen, C.-H.Sun, J.-Y.Bai, H.-R.Chen, and S.-Y. Chien, (2008) "Architectural analyses of K-means silicon intellectual property for image segmentation," in Proc. IEEE Int. Symp. Circuits Syst. (pp. 2578–2581).

[5] T.W. Chen and S.-Y.Chien,(2010) "Bandwidth adaptive hardware architecture of K-means clustering for video analysis," IEEE Trans. Very Large Scale Integr. (VLSI)Syst., vol. 18, no. 6, (pp. 957–966).

[6]Tung-Shouchen ,rong-changchen, (2005)"Acombined k-means and hierarchical clustering methodfor the clustering efficiency of microarray". (pp. 957–966).

[7] David casasemt and yu-chiang frank wang, (2006) "Hierarchical K-means Clustering Using New Support Vector Machines for Multi-class Classification". (pp. 2578–2581).

[8] Yi pan,R.Harrison, (2006) "Novel hybrid hierarchical k-means clustering method (HK-Means) for microarray analysis" (pp.355-368).

[9] Liwenchao,zhouyong, (2006) "Novel clustering algorithm based on hierarchical and k-means clustering", (pp.35-36).

[10]Shenfengchen and John H.Reif ,(2010) "Efficient lossless compression of trees and graphs", (pp.359-368).

[11] Jianjiun Ding, Wei-yi Wei, (2012) "Adaptive golomb coding for join geometrically distributed data and its application" (pp.55-68).

[12] ShanShan Yong, (2015)"Hardware implementation of real time lossless data compression and decompression Circuits", (pp.555-588).

[13] Akshatakoparde,Sandhiya, (2018) "Performance optimization of dual stage algorithm for lossless data compression and decompression", (pp.155-168).

[14] A. K. Jain, M. N. Murty, and P. J. Flynn, (1999)"Data clustering: A review," ACM Comput. Surveys, vol. 31, no. 3, (pp. 264–323).

[15] T.-S Chen, T.-H Tsai, Y.-T Chen, C.-C Lin, R.-C Chen, S.-Y Li, and H.-Y Chen (2005) "A combined K-means and hierarchical clustering method for improving the clustering efficiency of microarray," in Proc.IEEE Int. Symp. Intell. Signal Process. Commun. Syst., (pp. 405–408).

[16] N. Chehata and F. Bretar, (2008) "Terrain modeling from LIDAR data: HierarchicalK-means filtering and markovian regularization," in Proc. IEEE Int. Conf. Image Process. (pp. 1900–1903).

[17] D. Nistér and H. Stewénius, (2006) "Scalable recognition with a vocabulary tree," in Proc. IEEE Comput. Soc. Conf. Comput. Vision Pattern Recog., (pp. 2161 2168).

[18] J.Satheesh kumar,G.Saravana kumar (2018)" High performance decoding aware FPGA bit-stream compression using RG codes" Cluster Computing The Journal of Networks, Software Tools and Applications, (pp. 105–108).