A Study on Fast K-Means Clustering with Hierarchical Data Sampling for Image Processing

Tse-Wei Chen*, Daisuke Deguchi, Ichiro Ide, Hiroshi Murase (Nagoya University)

1. Introduction

Since many kinds of parallel architectures for video processing (1) are developed, the real-time processing is possible for image data with different sizes. However, the computational complexity of machine learning algorithms, which are key techniques for semantic analysis, is still high for resource-limited embedded systems. Therefore, the design of machine learning algorithms and hardware architectures turns into a challenging issue for the system development. Data clustering is an important part of machine learning algorithms, and K-Means is a well-known algorithm that has already been used in various kinds of applications. To accelerate the clustering speed of K-Means for image processing, a hierarchical data sampling technique is analyzed.

2. K-Means with Hierarchical Data Sampling

Data sampling is a technique that can be applied to K-Means clustering for speed acceleration (2). The centroids are initialized based on the K-Means clustering results of some samples of the whole data, and the clustering speed can be accelerated since fewer iteration numbers are needed when the positions of initialized centroids are close to those of final centroids. In this work, K-Means is executed with the hierarchical sampling technique, which means that the sampling technique is repeatedly applied to different hierarchical levels as shown in Fig. 1. When the level is 3, the sample rate is $1/2^3$, and when the level is 2, the sample rate is $1/2^2$. The hierarchical level is reduced sequentially and becomes 0 when all the input vectors are used. This method controls the sampling rate based on powers of two and is suitable for the implementation of hardware architectures. Different from previous works, the iteration number is not fixed in each level, and both the clustering speed and the clustering quality can be measured.

3. Experiments and Image Processing Applications

The hardware architecture for K-Means clustering based on the hierarchical sampling technique is implemented and verified. According to the hardware verification, Fig. 2(a) shows the analysis of the computational costs of K-Means with color clustering. It is observed that more than 50% of computation costs can be saved while the same clustering quality is maintained. Besides, a new application for K-Means, called edge-adaptive noise reduction, is also demonstrated. K-Means clustering is performed on the pixels in each window of an input image like a filtering process, and the cluster number is automatically estimated based on the Bayesian Information Criterion (BIC) (3). Fig. 2(b) and (c) show an example of the edge-adaptive filtering, which eliminates the noise and preserves the high-frequency details at the same time.

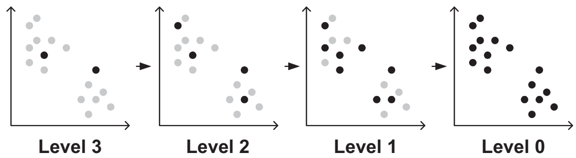

Fig. 1. Illustration of the technique of hierarchical data sampling. The black dots represent the data (vectors) that are used for K-Means in each hierarchical level.

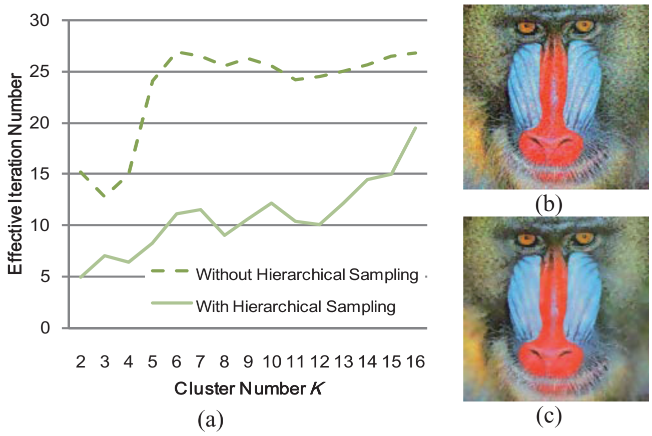

Fig. 2. (a) Analysis of clustering speed acceleration with color clustering. (b) Original image, “baboon,” which is corrupted with Gaussian noise. (c) Noise reduction result based on K-Means with cluster number estimation from $K = 1$ to $K = 4$.

4. Conclusions and Future Work

A study on K-Means clustering with the hierarchical data sampling technique, which can accelerate the clustering speed, is presented. The experiments show that the proposed method can be used for important image processing applications, such as color clustering and edge-adaptive noise reduction. For future developments, the acceleration and hardware implementation of other machine learning algorithms will be discussed.

References

(1) T. Kurafuji, *et al.*, “A scalable massively parallel processor for real-time image processing,” in ISSCC Dig. Tech. Papers, 2010, pp. 334–335.

(2) T. Saegusa and T. Maruyama, “An FPGA implementation of real-time K-Means clustering for color images,” J. Real-Time Image Processing, vol. 2(**4**), 309–318, Nov 2007.

(3) D. Pelleg and A. Moore, “X-Means: Extending K-Means with efficient estimation of the number of clusters,” in Proc. 7th Int. Conf. on Machine Learning, 2000, pp. 727–734.