1999년도 한국음향학회 학술발표대회 논문집 제18권 제2(s)호

수정된 K-means 알고리즘

김 형 철 , 조 재 황
동신대학교 전기전자공학과

Modified K-means algorithm

HyungCheol Kim , CheHwang Cho
Dept. of Electrical & Electronic Eng., Dongshin Univ.
khc519@yahoo.co.kr , chcho@dongshinu.ac.kr

Abstract

One of the typical methods to design a codebook is K-means algorithm. This algorithm has the drawbacks that converges to a locally optimal codebook and its performance is mainly decided by an initial codebook. D. Lee's method is almost same as the K-means algorithm except for a modification of a distance value. Those methods have a fixed distance value during all iterations. After many iterations, because the distance between new codevectors and old codevectors is much shorter than the distance in the early stage of iterations, the new codevectors are not affected by distance value. But new codevectors decided in the early stage of learning iterations are much affected by distance value. Therefore it is not appropriate to fix the distance value during all iterations. In this paper, we propose a new algorithm using each different distance value between codevectors for a limited iterations in the early stage of learning iteration. In the experiment, the result show that the proposed method can design better codebooks than the conventional K-means algorithms.

Ⅰ. 서 론

하나의 영상 안에서 추출된 n 차원 벡터들은 n 차원 벡터 공간 안에서 몇 개의 군집들(clusters)을 형성한다. 분할들의 표본으로 간주되는 군집들의 원형들은 코드북의 코드벡터들과 동일시된다. 즉, 코드북을 설계하는 것은 군집들의 원형들을 찾는 과정과 유사하다. 코드북을 설계하는 알고리즘 중에서 가장 대표적인 방법은 LBG(Linde, Buzo, and Gray) 알고리즘[1]이라고도 알려진 K-means 알고리즘이다. 이 알고리즘은 주어진 초기 코드북에 대하여 최소거리 조건과 중심조건을 이용하여 평균거리 오차가 최소가 되는 코드북을 반복조건에 따라 연속적으로 생성하는 것이다. 그러나 K-means 알고리즘은 국부적으로 최적화 되고, 그 성능이 초기 코드북에 크게 의존한다는 문제점을 가지고 있어 이를 보완하기 위해 K-means 알고리즘의 초기 코드북을 결정하는 많은 방법들이 제시되었는데[2]-[4], 그 중 splitting 방법이 다른 방법들보다 더 좋은 초기 코드북을 생성하는 것으로 알려져 있다.

K-means 알고리즘과 거의 동일하지만 각 반복과정에서 새로운 코드벡터를 구하는 방법만이 다른 알고리즘을 Jancey가 제안했는데[5], 이 방법은 그림 1에서와 같이 현재벡터와 새로운 군집의 중심점과 일직선상에 있는 반대편의 점, 즉 거리의 가중치(δ)가 2.0인 점을 새로운 코드벡터로 사용하지만 이 점이 수렴영역의 경계선에 놓여 임의의 데이터에 대하여 수렴이 되지 않는 경우가 있을 수 있다. 이러한 문제를 보완한 것이 D. Lee가 제안한 개선된 K-means 알고리즘이다[6]. D. Lee의 방법은 Jancey가 제안한 방법에서 현재벡터와 새로운 군집의 중심점과 일직선상에 있는 거리의 가중

치가 2.0인 점 대신, 거리의 가중치가 1.8인 점을 새로운 코드벡터로 사용하는 것으로 기존의 K-means 알고리즘보다 더 좋은 성능을 보인다. 그러나 코드벡터 생성 시 초기 반복학습상태에서는 새로운 코드벡터들이 거리의 가중치에 의해 많은 영향을 받으므로 모든 반복학습 과정 동안 가중치를 고정하는 것은 적절하지 않다. 따라서 본 논문에서는 기존 K-means방법과 Jancey, D. Lee의 방법에서 가중치를 고정하는 것과는 달리 반복학습 시 코드벡터간 거리의 가중치를 제한된 회수 동안 가변하는 방법을 제안한다. 제안한 방법에서는 제한된 반복학습 시까지 가중치를 각각 2.5, 3.0, 3.5로 변화하고, 많은 반복학습 과정 후에 생성된 새로운 코드벡터들은 가중치에 거의 영향을 받지 않으므로 제한된 반복학습 이후에는 가중치를 1.8로 고정한다.

II. 기존 K-means 알고리즘과 수정된 K-means 알고리즘

K-means 알고리즘은 두 가지의 조건, 즉 최소거리 조건과 중심조건을 만족하여야 한다. 최소거리 조건은 주어진 학습벡터와 코드벡터 사이의 Euclidean 거리가 최소일 때 학습벡터가 코드벡터에 대응하는 분할에 소속된다는 것을 말하며, 중심조건은 분할된 영역 안에서 학습벡터의 중심이 새로운 코드벡터가 된다는 것이다. K-means 알고리즘은 두 조건을 이용하여 평균거리 오차가 최소가 되는 코드북을 반복조건에 따라 연속적으로 생성하는 것으로 식 (1)과 같고,

$$y^i_{k+1} = y^i_k + \delta(m^i_{k+1} - y^i_k) \quad (1)$$

여기서, y^i_k 은 k 번 반복 시 i 번째 코드벡터, y^i_{k+1} 은 $k+1$ 번 반복 시 i 번째 코드벡터, m^i_{k+1} 은 $k+1$ 번 반복 시 i 번째 코드벡터에 대응되는 중심벡터이다.

그림 1에서 $\delta=1$인 경우 $y^i_{k+1} = m^i_{k+1}$으로 기존의 K-means 알고리즘을 나타내고, $\delta=2$ 인 경우는 Jancey의 방법, $\delta=1.8$ 인 경우는 D. Lee의 방법으로 위의 방법들 중에서 가장 좋은 성능의 코드북을 설계할 수 있다.

제안한 방법에서는 기존 K-means방법과 Jancey, D. Lee의 방법에서는 모든 반복학습과정 동안 거리의 가중치를 고정하였으나, 초기 반복학습 시에 가중치의 영향이 크다는 점을 이용하여 제한된 반복학습 시까지 가중치를 각각 2.5, 3.0, 3.5로 변화하고, 이후에는 가중치를 1.8로 고정하여 위의 방법들보다 더 좋은 성능의 코드북을 설계할 수 있다.

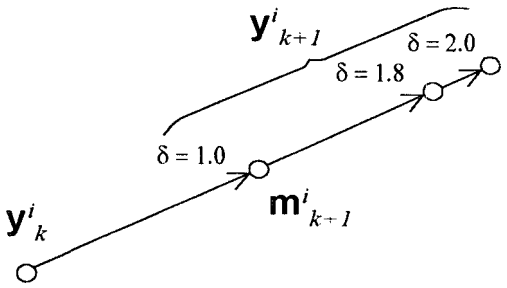

그림 1. 코드벡터와 중심벡터를 결정하는 거리의 가중치(δ)

III. 실험 및 결과

본 실험에서는 제안한 알고리즘과 기존 알고리즘을 비교하기 위해 256 그레이 레벨을 갖는 512×512 영상을 이용하여 16384개의 4×4 블록 단위로 블록킹한 후 이를 학습벡터로 사용하고, LENA 영상으로부터 splitting 방법을 사용하여 얻은 크기가 256인 코드북을 사용한다. 입력 영상은 LENA, PEPPERS, MANDRILL 영상을 이용하고, 원영상과 복원된 영상을 비교 평가하기 위한 RMSE(root mean square error)와 PSNR(peak signal to noise ratio)은 다음과 같다.

$$RMSE = \sqrt{\frac{1}{512^2}\sum_{i=1}^{512}\sum_{i=1}^{512}(f_{ij} - g_{ij})^2} \quad (2)$$

$$PSNR = 20\log_{10}\left(\frac{255}{RMSE}\right) \quad (3)$$

여기서 f_{ij}는 원영상의 화소값이고, g_{ij}는 복원된 영상의 화소값이다.

기존의 방법에서는 모든 반복학습과정 동안 거리의 가중치를 고정하여 초기 상태에서의 가중치의 영향을 고려하지 않았으나, 제안한 방법에서는 초기 반복학습 시에 가중치의 영향이 크다는 점을 이용하여 제한된 반복학습 시까지 가중치를 각각 2.5, 3.0, 3.5로 변화하고, 이후에는 가중치를 1.8로 고정하여 코드북을 설계한다.

그림 2, 그림 3, 그림 4는 LENA, PEPPERS, MANDRILL 영상에 대해 각각 다른 가중치에 따른 PSNR의 변화를 나타낸 것으로 그림 2에서 lena_25는 초기반복학습상태 즉, 1회 반복학습 시부터 순차적으로 7회 반복학습 시까지 가중치를 2.5로 하고, 이후 나머지 20회 반복학습 시까지는 가중치를 1.8로 한 경우로 5회 반복

시 까지 가중치를 2.5로 했을 때 PSNR이 가장 높게 나타난다.

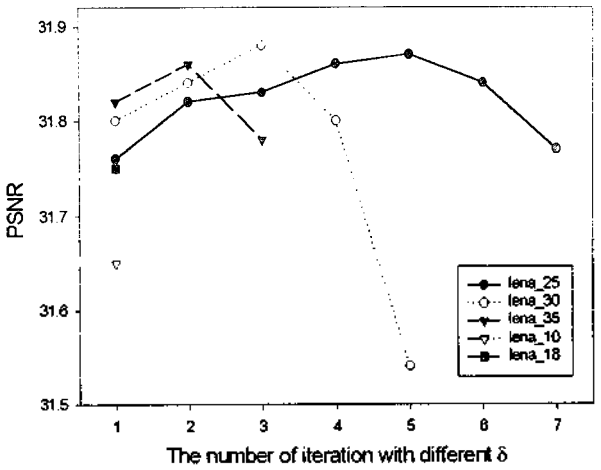

그림 2. LENA 영상의 PSNR

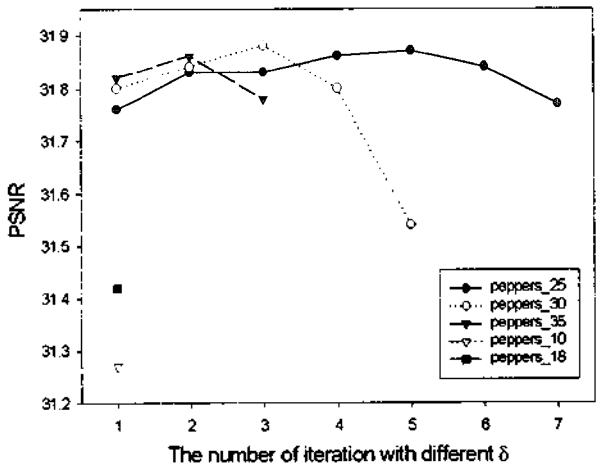

그림 3. PEPPERS 영상의 PSNR

lena_30은 초기반복학습 시의 가중치를 3.0으로 하고 이후 반복학습 시의 가중치를 1.8로 했을 경우로 3회 반복학습 시까지 가중치를 3.0으로 했을 때 가장 높은 PSNR을 보인다. lena_35는 초기반복학습 시의 가중치를 3.5로 하고 이후 반복학습 시의 가중치를 1.8로 했을 경우로 2회 반복학습 시까지 가중치를 3.5로 했을 때 가장 높은 PSNR을 얻을 수 있다. lena_10은 가중치가 1.0으로 고정된 기존의 K-means 알고리즘이며, lena_18은 D. Lee의 방법을 나타낸다.

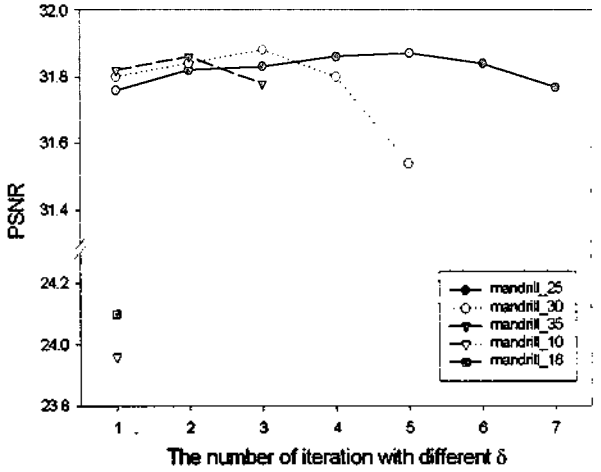

그림 4. MANDRILL 영상의 PSNR

실험 결과 가중치가 증가할수록 더 적은 반복회수에서 가장 높은 PSNR을 얻을 수 있고, 따라서 초기반복학습 시 가중치의 영향이 크다는 것을 확인할 수 있다. 기존의 두 방법에서는 각 영상들마다 PSNR이 크게 차이가 나타나는 것을 볼 수 있으나 제안된 방법에서는 3개의 입력 영상에 대해 PSNR이 거의 동일하게 분포함을 알 수 있다.

(a) Original LENA 영상

(b) 복원된 LENA 영상
(PSNR=31.88dB)

그림 5. 원영상과 복원된 LENA 영상

Ⅳ. 결 론

코드벡터 생성 시 새로운 코드벡터들이 거리의 가중치에 의해 많은 영향을 받는다. 기존의 K-means 알고리즘, Jancey의 방법 그리고 D. Lee의 방법에서는 모든 반복학습과정 동안 거리의 가중치를 고정하여 코드북을 설계하였다. 그러나 코드벡터 생성 시 초기 반복학습상태에서는 새로운 코드벡터들이 거리의 가중치에 의해 많은 영향을 받고, 많은 반복학습 과정 후에 생성된 새로운 코드벡터들은 가중치에 거의 영향을 받지 않으므로 모든 반복학습과정 동안 가중치를 고정하는 것은 적절하지 않다. 따라서 제안한 방법에서는 반복학습 시 코드벡터간 거리의 가중치를 제한된 회수 동안 각각 2.5, 3.0, 3.5로 가변하고 이후 반복학습 시의 가중치를 1.8로 고정하였다. 그 결과 기존의 방법보다 코드북의 성능을 향상할 수 있었고, 초기반복학습 시 가중치의 영향이 크다는 것을 알 수 있다.

참고문헌

1. Y.Linde, A.Buzo, and R.M. Gray, "An algorithm for vector quantizer design", IEEE Trans. Commun., vol. COM-28, pp. 84-95, 1980.
2. W. H. Equitz, "A new vector quantization clustering algorithm", IEEE Trans. Acoust. Speech and signal Proc., pp. 1568-1575, 1989.
3. I.Katsavounidis, C.C. Jay Kuo, and Z.Zhang, "A new initialization technique for generalized Lloyd iteration", IEEE Signal Processing Letters, vol. 1, pp. 144-146, 1994.
4. M.Rabbani and P.W. Jones, *Digital image compression techniques*, SPIE Press, 1991.
5. M.R. Anderberg, *Cluster analysis for applications*, Academic, New York, 1973.
6. D.Lee, S.Baek, and K.Sung, "Modified K-means algorithm for vector quantizer design", IEEE Signal Processing Letters, vol. 4, pp. 2-4, 1997.