초기 반복학습 시 수렴영역을 벗어난 가중치에 의한 K-means 알고리즘

박 소 희 , 조 제 황
동신대학교 전기전자공학부

K-means Algorithm in outside weight region of convergence for initial iteration learning

SoHee Park , CheHwang Cho
Dept. of Electrical & Electronic Eng., Dongshin Univ.
hee023@hanmail.net , chcho@white.dongshinu.ac.kr

요 약

본 논문에서는 랜덤초기화 방법을 사용하여 초기 코드북을 생성하고, 이를 이용하여 초기 반복학습 시 수렴영역을 벗어난 2 이상의 가중치에 의한 K-means 알고리즘을 제안한다.

기존의 K-means 알고리즘이 국부적으로 최적화 되고 초기 반복학습 시에 가중치의 영향이 크다는 점을 이용하여, 제안된 방법에서는 초기 반복학습 시의 가중치를 수렴영역에서 벗어난 큰 값으로 주고 이후 반복학습시의 가중치는 수렴영역 안에 있는 값으로 고정하여 코드북을 설계한다. 또한 초기 코드북을 얻기 위해 Splitting 방법과 같은 추가적인 과정 없이 랜덤한 방법에 의한 초기 코드북을 적용함으로써 제안된 알고리즘이 단순한 구조를 가지며, 구해진 코드북의 성능도 우수함을 확인할 수 있었다.

I. 서 론

멀티미디어의 발달로 인하여 다양한 형태의 영상과 음성 정보를 활용하게 됨으로써 처리해야할 정보의 양이 증가하게 되었다. 멀티미디어 정보 중에서 가장 많은 데이터량을 필요로 하는 영상은 실시간적 처리나 저장매체의 대용량화 같은 어려움이 따른다. 따라서 영상의 품질을 원하는 수준으로 유지하며 영상의 데이터량을 줄일 수 있는 압축기술이 중요시된다. 이러한 영상 압축은 크게 무손실 압축과 손실 압축으로 나누어지는데, 본 논문에서는 손실 압축 방법중의 하나인 벡터 양자화(Vector Quantization) 방법을 사용한다[1].

벡터 양자화 방법은 데이터원이 통계적으로 독립된 심볼로 구성되더라도 스칼라 대신 벡터로 조합된 신호를 부호화 함으로써 높은 압축률을 얻을 수 있다는 Shannon의 왜곡률이론(rate-distortion theory)에 근거를 두고 있다. 벡터 양자화 방법은 영상을 몇 개의 성분을 갖는 입력벡터로 전처리하고, 학습 알고리즘에 의해 얻어진 코드벡터로 구성된 코드북으로부터 각 입력벡터에 대해 가장 근사한 코드벡터를 찾아 그 벡터에 부여된 색인(index)을 찾는 방법이다.

코드북을 설계하는 알고리즘 중에서 가장 대표적인 방법은 LBG(Linde, Buzo, and Gray) 알고리즘이라고도 알려진 K-means 알고리즘이다[2]. 이 알고리즘은 주어진 초기 코드북에 대하여 최소거리 조건과 중심조건을 이용하여 평균거리 오차가 최소가 되도록 반복 학습하여 코드북을 생성하며 이때 사용되는 초기 코드북을 결정하는 여러 가지 방법들이 제시되었다[3]-[6]. 그러한 방법들 중에서 Splitting 방법이 다른 방법들보다 더 좋은 초기 코드북을 생성하는 것으로 알려져 있다.

K-means 알고리즘과 거의 동일하지만 각 반복과정에서 중심 조건이 변형된 알고리즘을 Jancey가 제안했

는데[7], 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 경계선이 놓여 임의의 데이터에 대하여 수렴이 되지 않는 경우가 있을 수 있다. 이러한 문제를 보완한 것이 D. Lee가 제안한 개선된 K-means 알고리즘이다[8]. D. Lee의 방법은 현재벡터와 새로운 군집의 중심점과 일직선상에서 거리의 가중치가 2.0인 점 대신 거리의 가중치가 1.8인 점을 새로운 코드벡터로 사용하는 것으로 기존의 K-means 알고리즘 보다 더 좋은 성능을 보인다. 하지만 이러한 K-means 알고리즘들은 국부적으로 최적화 되며, 그 성능이 초기 코드북에 크게 의존한다는 문제점을 가지고 있다.

본 논문에서는 초기 코드북을 얻기 위해 Splitting 방법과 같은 추가적인 과정 없이 랜덤한 방법에 의해 간단히 초기 코드북을 얻으며, 초기 반복학습 시의 가중치를 수렴영역에서 벗어난 큰 값으로 주고 이후 반복학습시의 가중치는 수렴영역 안에 있는 값으로 고정하여 코드북을 설계하는 개선된 K-means 알고리즘을 제안한다.

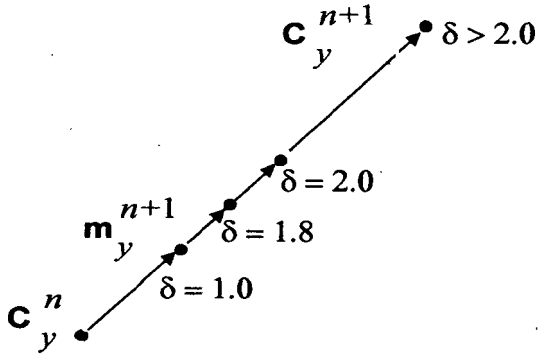

그림 1. 거리의 가중치(δ)에 따른
K-means 알고리즘

II. 제안된 알고리즘

N차원의 유클리드 공간 R^N에서 K크기의 입력벡터의 집합을 I라 하면 $I=\{ i_1, i_2, \cdots, i_K \}$이 되며, I에서 임으로 L개의 입력벡터를 선택한 경우의 집합 $C=\{ c_1, c_2, \cdots, c_L \}$는 크기 L의 초기 코드북이 된다. 코드북은 K개의 입력벡터를 C 군집에 할당함으로써 설계되며 이러한 코드북 설계의 평가는 다음과 같은 평균 왜곡에 의해 측정된다.

$$D = \frac{1}{K} \sum_{x=1}^{K} d_{\min}(i_x) = \frac{1}{K} \sum_{x=1}^{K} \min_{c_y \in C} d(i_x, c_y) \quad (1)$$

K-means 알고리즘은 최소거리조건에 근거하여 각 입력벡터를 어떤 한 군집에 할당한다. 이 알고리즘에 의해 $d(i_x, c_y) = d_{\min}(i_x) = \min_{c_y \in C} d(i_x, c_y)$이면 입력벡터 i_x는 y번째 군집에 할당된다. 여기서, $d(i_x, c_y)$는 $\| i_x - c_y \|^2$로 정의되는 입력벡터 i_x와 코드벡터 c_y의 유클리드 거리의 제곱이다. 또한 최소거리 조건은 다음과 같이 정의되는 선택함수로 표현될 수 있다.

$$p_y(i_x) = \begin{cases} 1 & d(i_x, c_y) = d_{\min}(i_x) \text{ 일때} \\ 0 & \text{위조건이아닐때} \end{cases} \quad (2)$$

K-means 알고리즘은 최소거리조건에 근거하여 각 학습벡터를 단일 군집에 할당하는데, 이 경우에 crisp 결정에 근거하여 0 아니면 1을 할당한다[9]. 코드벡터는 다음과 같이 정의되는 왜곡측정을 최소화하여 얻어진다.

$$J = \sum_{y=1}^{L} \sum_{x=1}^{K} p_y(i_x) \| i_x - c_y \|^2 \quad (3)$$

주어진 일련의 소속함수에 대한 코드벡터는 다음과 같이 $J = J(c_y, y = 1, 2, \cdots, L)$을 최소화함으로써 평가될 수 있다.

$$c_y = \frac{\sum_{x=1}^{K} p_y(i_x)\, i_x}{\sum_{x=1}^{K} p_y(i_x)} \quad \forall y = 1, 2, \cdots, L \quad (4)$$

중심조건에 의해 유클리드 중심 혹은 y번째 군집에 할당된 모든 입력벡터의 중심이 코드벡터 c_y가 된다. 위의 식에서 새로운 코드벡터를 구하기 위해 현재 코드벡터와 새로운 군집의 중심점과 일직선상에 있는 거리의 가중치를 적용하면 다음과 같다.

$$c_y^{n+1} = c_y^n + \delta(m_y^{n+1} - c_y^n) \quad (5)$$

여기서, c_y^n은 n번 반복 시 코드벡터, c_y^{n+1}은 $n+1$번 반복 시 코드벡터, m_y^{n+1}은 $n+1$번 반복 시 코드벡터에 대응되는 중심벡터이다.

식 (5)에서 $n=1$일 때, 즉 초기 반복학습일 때 거리의 가중치 δ를 수렴영역에서 벗어나게 2보다 큰 값을 주고 그 이후에는 가중치를 고정한다.

Ⅲ. 실험 및 결과

본 실험에서는 제안한 알고리즘과 기존 알고리즘을 비교하기 위해 256 그레이 레벨을 갖는 512×512 크기의 LENA, PEPPER 영상을 4×4 블록단위로 나누어 입력벡터로 사용한다. 또한 기존 알고리즘에 사용될 초기 코드북은 Splitting 방법에 의해 생성된 256 크기의 코드북을 사용한다. 원 영상과 복원된 영상을 비교 평가하기 위한 PSNR(Peak to Signal Noise Ratio)은 다음과 같다.

$$PSNR = 20 \log_{10}\left(\frac{255}{\sqrt{\frac{1}{512^2}\sum_{i=1}^{512}\sum_{i=1}^{512}(f_{ij}-g_{ij})^2}}\right) \quad (7)$$

여기서 f_{ij}는 원 영상의 화소 값이고, g_{ij}는 복원된 영상의 화소 값이다.

첫 번째 실험은 LENA와 PEPPER 영상에서 초기 가중치를 3.0~5.0로 주고 2번째 반복학습부터는 1.0으로 고정시킨 다음, 20회 반복학습 후의 PSNR을 기존의 K-means 알고리즘과 비교하였다.

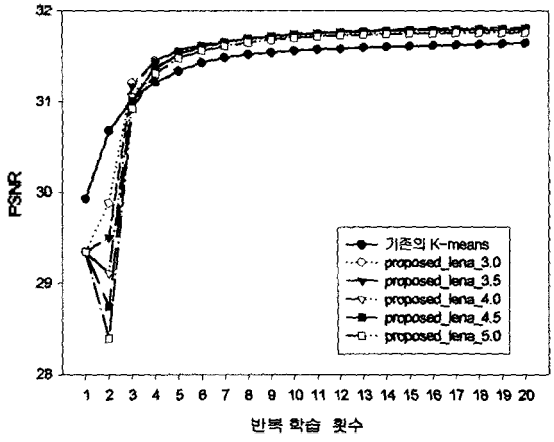

그림 2. LENA 영상의 반복학습 횟수에 따른 PSNR(δ =1.0)

그림 2와 그림 3은 랜덤 초기 코드북을 사용하는 제안된 알고리즘과 Splitting 방법의 초기 코드북을 사용하는 기존의 K-means 알고리즘의 PSNR을 비교한 결과이다. 초기 반복 시에는 기존의 방법이 우수하나 20번 반복학습 후에는 제안된 방법이 더 우수함을 알 수 있다.

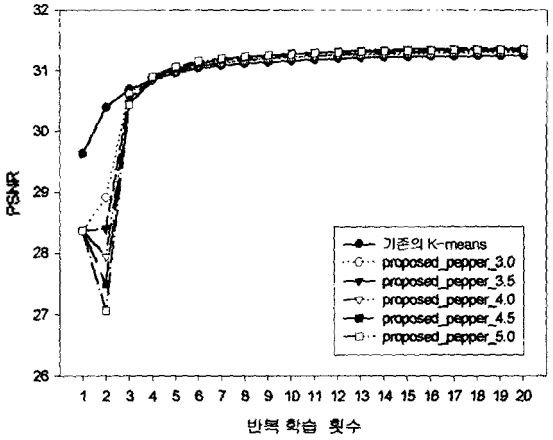

그림 3. PEPPER 영상의 반복학습 횟수에 따른 PSNR(δ =1.0)

두 번째 실험은 LENA와 PEPPER 영상에서 초기 가중치를 3.0~5.0로 주고 2번째 반복부터는 1.8로 고정시킨 다음, 20회 반복학습 후의 PSNR을 기존의 D. Lee가 제안한 K-means 알고리즘과 비교하였다.

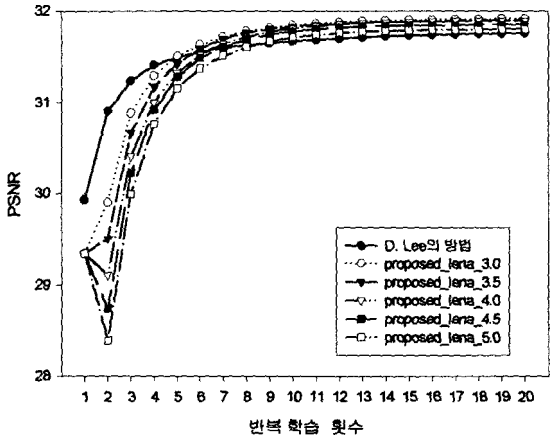

그림 4. LENA 영상의 반복학습 횟수에 따른 PSNR(δ =1.8)

그림 4와 그림 5는 랜덤 초기 코드북을 사용하는 제안된 알고리즘과 Splitting 방법의 초기 코드북을 사용하는 D. Lee의 알고리즘과의 PSNR을 비교한 결과이다. 마찬가지로 초기 반복 시에는 기존의 방법이 우수하나 20번 반복학습 후에는 제안된 방법이 더 우수함을 알 수 있다.

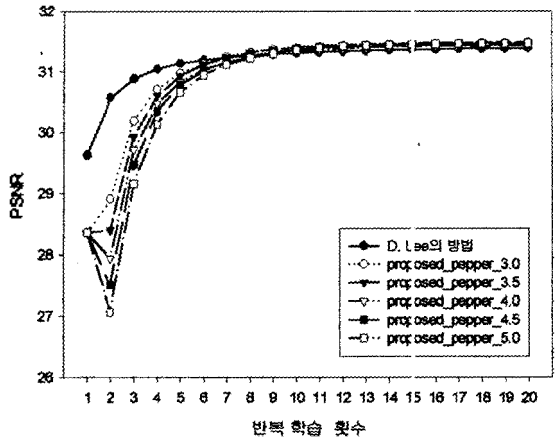

그림 5. PEPPER 영상의 반복학습 횟수에 따른 PSNR(δ=1.8)

따라서, 그림 2, 3, 4, 5의 결과처럼 랜덤 초기 코드벡터를 사용하며 첫 반복학습에서 수렴영역을 벗어난 3.0~5.0사이의 가중치를 주고 나머지 반복학습에서는 1.0 혹은 1.8로 고정시키는 제안된 알고리즘이 기존방법에 비해 우수한 코드북 설계가 가능함을 알 수 있다.

Ⅳ. 결 론

본 논문에서는 초기 코드북으로 랜덤 초기 코드벡터를 가지며 초기 반복학습 시의 가중치를 수렴영역에서 벗어난 큰 값으로 주고 이후 반복학습시의 가중치는 수렴영역 안에 있는 값으로 고정시키는 새로운 코드북 생성 알고리즘을 제안하였다. 제안된 방법을 기존의 K-means 알고리즘과 D. Lee가 제안한 방법에 대해서 모두 실험한 결과 제안한 방법에서 초기 반복학습 시의 가중치가 3.0~5.0 일 때 제안된 방법의 성능이 기존 방법보다 우수함을 알 수 있다.

참고문헌

[1] T.Murakami, K.Asai and E.Yamazaki, "Vector quantizer of video signals", Electronics Letters, vol.7, pp. 1005-1006, 1982

[2] Y.Linde, A.Buzo, and R.M.Gray, "An algorithm for vector quantizer design", IEEE Trans. Commun., vol. COM-28, pp. 84-95, 1980.

[3] R.M.Gray, "Vector quantization", IEEE ASSP Mag., pp. 4-29, 1984

[4] W.H.Equitz, "A new vector quantization clustering algorithm", IEEE Trans. Acoust Speech and signal Proc., pp. 1568-1575, 1989.

[5] 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.

[6] M.Rabbani and P.W. Jones, *Digital image compression techniques*, SPIE Press, 1991.

[7] M.R. Anderberg, *Cluster analysis for applications*, Academic, New York, 1973.

[8] 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.

[9] M. Friedman and A. Kandel, *Introduction to pattern recognition*, World Scientific, 1997.