論 文

論文 75-12-2-1

寫眞 데이타를 爲한 한 Adaptive Data Compression 方法

(An Adaptive Data Compression Algorithm for Video Data)

金 在 均*
(Kim, Jae-kyoon)

要 約

本 論文은 寫眞데이타에 쉽게 適用할수있는 한 adaptive data compression 方法을 提示하였다. 이웃 sample data 사이의 높은 correlation때문에 발생할 符號化 複雜性을 간편한 sample difference data로 對處하였으며, data의 statistical nonstationarity에 適應키 위해서 여덟가지 符號(code)로 구성된 code set중에서 最適符號를 選擇토록 하였다. code set는 두가지 等長符號와 여섯가지 修正型 Shannon-Fanco 符號로 하였다. difference data의 確率分布는 Laplacian model로, entropy의 確率分布는 Gaussian model로 하였다. 符號選別 parameter로서 entropy와 p_r〔差異값=0〕$\equiv p_0$를 比較하였다.

콤퓨타 實驗結果 이 adaptive coding 方法으로 2對 1의 데이타 減縮比를 얻었다. 이 方法은 fixed coding에 比해서 데이타 減縮比와 符號化效率에서 約 10%와 15%의 利得을 주었다. 또한 p_0는 entropy보다 훨씬 간편한 符號選別 parameter로 동시에 entropy 경우와 1% 內外의 좋은 結果를 얻을 수 있음이 確認되었다.

Abstract

This paper presents an adaptive data compression algorithm for video data. The coding complexity due to the high correlation in the given data sequence is alleviated by coding the difference data sequence rather than the data sequence itself. The adaptation to the nonstationary statistics of the data is confined within a code set, which consists of two constant length codes and six modified Shannon-Fano codes. It is assumed that the probability distributions of the difference data sequence and of the data entropy are Laplacian and Gaussion, respectively. The adaptive coding performance is compared for two code selection criteria: entropy and p_r〔difference value=0〕$\equiv p_0$.

It is shown that data compression ratio 2 : 1 is achievable with the adaptive coding. The gain by the adaptive coding over the fixed coding is shown to be about 10% in compression ratio and 15% in code efficiency. In addition, p_0 is found to be not only a convenient criterion for code selection, but also such efficient a parameter as to perform almost like entropy.

* 正會員, 韓國科學院 電氣 및 電子工學科.
本 硏究는 韓國科學院 自體硏究費와 美國 NRC Research Associateship의 支援을 받았음.
The author is with the Department of Electrical Engineering, the Korea Advanced Institute of Science (KAIS), Seoul, Korea 131. This work was supported in part by the KAIS Internal Research Fund and in part by the U.S. National Research Council Research Associateship.
接受日字 : 1975年 2月 1日

1. 序 論

데이타 減縮(data compression)方法을 생각함에 있어서 필요한 前提條件의 하나는 情報損失 즉 歪曲(distortion)의 허용 범위이다. 주어진 데이타가 이미 analog signal을 量子化(quantize)해서 얻은 digital signal인 경우, 이 데이타는 처음부터 量子化誤差

(quantization error)를 포함하고 있다. 그러나 本論文에서는 이 주어진 digital data sequence를 더 이상의 誤差없이 좀더 짧게 변환하려고 한다(inversible transformation). 이 變換過程에서 제거된 데이타의 冗長分(redundancy)은 受信側에서 再生하므로 原來의 data sequence를 다시 얻을 수 있게 된다. 주어진 digital data를 誤差 增加없이 減縮하기 위한 方法은 결국 能率的인 source coding 法이다. 따라서 本 論文에서는 주어진 데이타源의 統計的 特性에 적합한 한 adaptive coding 방법을 紹介하고저 한다.

實際 데이타源은 간편한 memoryless stationary sequence가 아닌 복잡한 nonstationary data sequence를 發生하며 이웃 데이타 間에는 높은 correlation(相關性)이 있다. 이 correlation을 n次의 Markov source로 近似的으로 해석할 수도 있겠으나 實際的인 符號化(coding)가 어렵다. 즉 符號化에 필요한 $(n+1)$次의 joint *PDF*(probability density function)을 算出 維持하기가 real-time 應用에서는 너무 복잡한 일이다. 시간에 따라서 變化하거나 잘 알수 없는 確率特性을 갖인 信號에 대해서도 adaptive coding 방법이 해석적으로 가능하기는 하나,[1)2)] 대체로 긴 block coding에 따르는 복잡성 문제가 있다. 실제의 응용에서는 데이타 相互間 및 data block 相互間이 統計的으로 獨立이라는 近似的 前提아래에서, 미리 만들어둔 몇개의 符號(code)* 가운데서 주어진 確率特性에 가장 適合한 符號를 선택하는 방법이 있다.[3)] 이로서 變化하는 統計的 性質에 쉽게 適應(adaptation)하는 符號를 찾을 수 있어 全體的으로 높은 符號化效率을 얻을 수 있게 된다. 여기서는 물론 주어진 데이타로 부터 相互間 統計的 獨立에 가까운 data sequence를 먼저 만들 必要가 있다.

本 論文에서는 origenal data sequence 代身 data difference sequence를 符號化하므로서 높은correlation에 따른 符號化複雜性을 避했으며, nonstationary인 統計적 特性에는 여덟가지의 符號(code)로서 適應토록 하였다. 한가지 符號만을 이용한 nonadaptive 符號化는 이미 real-time으로 응용이 되고 있다.[4)] 이제 符號化에서의 基本이 되는 entropy와 correlation, 데이타의 統計的 特性, 符號(code) 및 符號選別을 중심으로 考察코저 한다.

* "符號"는 code 보다도 codeword나 code alphabet로 誤解될 가능성이 많아서 필요할 때마다 그 眞意를 英字로 明記하였음. 一般化되지 않은 用語는 처음 使用될 때 英字를 附記하였음.

2. Entropy와 Correlation

本 論文에서 考察할 데이타源은 ATS-3(Applications Technology satellite) 인공위성에서 찍은 黑白사진으로서 기상예보를 위한 지구구름 사진이다(그림 1).

그림 1 ATS—3 사진데이타
Fig.1 ATS—3 Picture data

사진은 2400개의 走査線(scan line)으로, 各 走査線은 4096 sample(picture element)로 되어 있으며 各 sample은 여섯 bit로 均一하게 量子化(uniform quantization)되어 있는 데이타이다. 따라서 이 사진 데이타를 그대로 傳送한다면 즉 PCM 傳送인 경우 약 5898萬 bit를 보내야 한다. 사진에서 알 수 있는 바와 같이 네 귀의 검은 공간에 해당하는 데이타만을 식별해도 상당한 데이타 減縮을 얻을 수 있다. 지역에 따라서 明暗이 크게 달라 data sequence가 nonstationary임을 곧 알수 있으며, 또한 한 走査線 上에서도 이웃 data 間의 큰 類似性으로 부터 높은 correlation을 짐작할 수 있다.

여기서 주어진 sample sequence를 $\{s_n\}$이라 하면 各 s_n은 여섯 bit로 量子化된 것이므로 64가지의 明暗 level中 한가지 값을 갖는 random variable(確率變數或은 無作爲變數)이라 할 수 있다. 여기서 s_n 自體의 *PDF*도 중요하지만 source coding에서 모든 比較의 尺度가 되는 entropy를 먼저 생각해 볼 필요가 있다. 이는 Shannon의 source coding 定理에 의해서 符號化로서 얻을 수 있는 평균 符號(code word) 길이의 下限은 바로 이 데이타源의 entropy이기 때문이다.

stationary random sequence $\{x_i\}$를 만들어 내는 한

情報源 X의 entropy $H(X)$는 다음과 같이 定義된다.[5]

$$H(X)=\lim_{n\to\infty}G_n(X)\equiv\lim_{n\to\infty}\frac{1}{n+1}H(x_0,x_1,\cdots,x_n)$$
$$=\lim_{n\to\infty}H_n(X)\equiv\lim_{n\to\infty}H(x_n|x_{n-1},x_{n-2},\cdots,x_0) \quad (1)$$

여기서 $H_0(X)=G_0(X)=H(x_0)$, $G_n(X)\geq 0$, $H_0(X)\geq 0$ 이며, joint entropy $G_n(X)$와 conditional entropy $H_n(X)$는 모두 n가 增加함에 따라서 monotonically 減少하므로 한 極限값(limit) 즉 바로 entropy값 $H(X)$에 收斂한다. 이 收斂速度가 늦는 경우 즉 이웃 data 間의 correlation이 큰 경우에는 entropy 測定이 어렵다. 그러나 correlation이 작은 경우에는 이 측정이 비교적 쉽게되며, memoryless source에서는 $H(X)=H(x_0)$ 가 되어 간단히 구할 수도 있다.

이제 주어진 source data sequence $\{s_n\}$을 1對 1 mapping으로 다른 sequence $\{d_n\}$으로 變換했다면

$$\{s_n\}\leftrightarrow\{d_n\} \quad (2)$$

$\{s_n\}$와 $\{d_n\}$ 사이에는 아무런 情報上의 變化가 없다. 따라서 두 random sequence의 平均 情報인 entropy도 같아야 한다.

$$H(S)=H(D) \quad (3)$$

여기서 만약 새로운 sequence $\{d_n\}$이 memoryless source에 近似할 만큼 correlation이 작다면 $H(D)$의 測定은 물론이고 符號化도 크게 簡素化될 수 있다. 또한 $\{s_n\}$을 $\{d_n\}$으로 變換하는 過程에서 일바간의 誤差를 感受한다면 $H(D)<H(S)$로 만들 수 있게되어 보다 큰 데이타 減縮을 얻을 수 있는 所地를 마련할 수도 있다.

이웃 데이타 사이의 correlation을 減少시킨다는 것은 곧 이웃 데이타間의 冗長度를 줄인다는 것이라 볼 수 있다. 冗長度가 큰 경우 즉 correlation이 높은 경우에는 주어진 data sequence $\{s_n\}$의 進行過程을 쉽게 미리 豫測할 수 있다. 이때 豫測할 수 있는 量은 곧 우리가 여기서 除去해 보려는 冗長分에 相應하는 量이라고 볼 수 있다. 따라서 주어진 데이타 즉 實際값(real value) s_n과 이를 豫測한 값(predicted value) $\hat{s}_n$과의 差異값(difference value) d_n으로 된 새로운 sequence

$$\{d_n\}=\{s_n-\hat{s}_n\} \quad (4)$$

은 original data sequence $\{s_n\}$보다 correlation이 작게 된다. 이를테면 $\{s_n\}$이 wide-sense Markov stationary process인 경우 s_n을 最少平均自乘法(Least mean square error)으로 linear predict 하면 $\{d_n\}$은 uncorrelated sequence가 된다.[6] 또한 平均自乘値 $E[d_n^2]$을 最少化함은 바로 d_n의 變位(deviation)를 減少시키는 結果가 되므로 1st order entropy $H(d_0)$가 減少됨은 물론이다.

여기서 고찰하려는 사진 데이타와 같이 correlation이 아주 높은 경우는 豫測값 $\hat{s}_n$을 s_n 바로 直前값으로 해도 즉

$$\{d_n\}=\{s_n-s_{n-1}\} \quad (5)$$

으로 해도 여러개의 data 값으로 부터 求하는 경우 즉 $s_n=f(s_{n-1},s_{n-2},\cdots,s_{n-N})$인 경우와 거의 같은 結果를 얻을 수 있다. 이렇게 되면 $\{d_n\}$을 쉽게 얻을 수 있음은 물론이고 相關係數(correlation coefficient) $\rho_d(i)$도 다음과 같이 크게 줄어든다(부록 참조).

$$\rho_d(i)\cong\begin{cases}-0.5, & i=\pm 1\\ 0, & i\neq 0,\pm 1\end{cases} \quad (6)$$

따라서 이 $\{d_n\}$을 實際應用面에서는 거의 uncorrelated data sequence로 볼 수도 있다.

3. 統計的 特性

前節에서는 주어진 source data sequence $\{s_n\}$과 꼭 같은 情報을 갖이고 있으면서도 correlation이 매우 작은 (5)식의 새로운 data sequence $\{d_n\}$의 entropy와 *PDF*를 考察하면 충분하다는 결론을 얻었다. 그리고 序論에서 언급된바와 같이 *PDF* 算出, 符號化의 簡便性등 실제응용상의 便宜를 위해서 $\{d_n\}$을 統計的으로 獨立한 (statistically independent) data sequence라고 보겠다.

그러나 데이타源의 nonstationarity는 如前히 $\{d_n\}$에도 남아있다. $\{d_n\}$의 통계적 특성은 변화하는 반면에, 제한된 個數 즉 여덟가지 符號(code)로서 adaptive coding을 한다면 항상 $\{d_n\}$의 *PDF*에 맞는 符號를 선택하기는 實로 不可能한 일이다. 따라서 이런 경우 *PDF*를 實測한 後에 이에 맞는 code를 만들기 보다는 典型的인 *PDF* 模型을 活用하는 것이 오히려 유리하다고 할 수 있다. 其實 사진 데이타에서 difference sequence $\{d_n\}$은 Laplacian 分布(double exponential)에 매우 近似하다는 것이 알려져 있다.[7] 물론 이런 近似化의 正確度는 사진특성, sampling rate, 量子化 level등 여러 가지에 달려 있기는 하다.

3—1. Laplacian 分布와 Entropy

豫測値 s_n을 여러개의 데이타값으로 부터 求한다면 s_n은 연속적인 確率變數가 된다. 따라서 (4)식에 의해서 差異값도 連續的인 確率變數가 된다. 이 差異값이 連續的인 Laplacian 分布 $f(x)$를 갖는다면, 均一하게 量子化된 d_n의 確率分布는 다음과 같다.

$$P[d_n=i]=\int_{i-0.5}^{i+0.5}f(x)dx=\int_{i-0.5}^{i+0.5}\alpha e^{-2|\alpha|x}dx$$
$$=\begin{cases}1-e^{-\alpha}, & i=0\\ e^{-2\alpha|i|}\sinh\alpha, & i=\pm 1,\pm 2,\cdots\end{cases} \quad (7)$$

여기서 $\alpha>0$ 이며, 積分범위가 ± 0.5로 제한된 것은 差

異값을 量子化함에 있어서 더 以上의 量子化誤差를 許容치 않으려는 기본 要件 때문이다. 물론 量子化間隔에는 제한이 있으므로 確率分布의 模型을 찾는 過程에서는 (7)식의 적분범위를 좀더 융통성 있게 조절할 수도 있는 일이다.

$\{s_n\}$의 變位범위가 〔0, 63〕이므로 (7)식에서 i의 값은 $i=\pm 63$까지 생각할 수 있다. 그러나 實際 符號(code)를 구성할 때 code dictionary의 크기를 제한하기 위해서 i의 값이 一定値 以上일 때는 긴 符號(codeword)를 使用하는 대신에 데이타값 s_n 自體를 보내는 것이 더욱 便利하다. 個別的인 符號(codeword)를 配定할 i의 上限값 M를 다음과 같이 정할 수 있다.

$$M=\min_{j}\{〔1-\sum_{|i|\geq j} p〔i〕〕\leq\delta\} \qquad (8)$$

여기서 $p〔i〕=p〔d_n=i〕$이며, δ(=tail probability)는 便宜上 entropy가 $H(D)\leq 3.5$인 경우에는 0.001, $H(D)>3.5$인 경우에는 0.01로 취했다. 여기서 δ의 값이 매우 작으므로 entropy 계산에서 δ가 한 量(entity)로 들어가도 別差異 없음을 알 수 있다. (7), (8)식으로부터 Laplacian entropy $H(\alpha)$는 다음과 같다.

$$H(\alpha)=-\sum_{i=-M}^{M} p〔i〕\log_2 p〔i〕-\delta\log_2\delta \qquad (9)$$

이 $H(\alpha)$의 곡선은 그림 2와 같다. 이 그림에서 $\alpha\leq 1.28$인 경우에는 $H(\alpha)$가 다음과 같이 近似式으로 表現됨을 알 수 있다.

$$H_a(\alpha)=1.56-3.16\log_{10}\alpha, \quad \alpha\leq 1.28 \qquad (10)$$

이 $H_a(\alpha)$는 그림 2에 있는 $P〔d_n=0〕\equiv P_0$ 곡선과 더불어 符號(code) 選別에 매우 편리한 關係式이다. 이는 $\{d_n\}$의 確率分布가 과연 Laplacian이라면 code 選別過程에서 確率分布, entropy $P〔d_n=0〕$이 모두 같은 결과를 주는 基準(criterion)이 될 수 있기 때문이다. 그러나 $P〔d_n=0〕$을 使用하면 選別過程이 훨씬 편리하고 신속하다는 장점이 있다.

3—2. 測定 Entropy의 分布

그림 1의 사진 데이타에서 求한 走査線分(scan line segment) 別 測定 entropy H_s와, 測定된 $P〔d_n=0〕$에서 얻은 Laplacian 近似 線分 entropy H_{sa}의 分布는 그림 3과 같다. 여기서 H_{sa}는 (7), (10)에 의해서 다음과 같이 얻은 값이다.

$$H_{sa}=H_a(P〔d_n=0〕)$$
$$=1.56-3.16\log_{10}(-\log_e(1-P〔d_n=0〕)) \qquad (11)$$

그림 2에서 볼 수 있는 바와 같이 이 式은 대체로 $H_s\geq 1.25$인 경우에만 有效한 式이다. 그러나 그림 3에서와 같이 $0.1\leq H_s\leq 2$인 범위에서 H_s는 거의 均一 分布를 하고 있으므로 (11)式의 誤差는 거의 문제되지

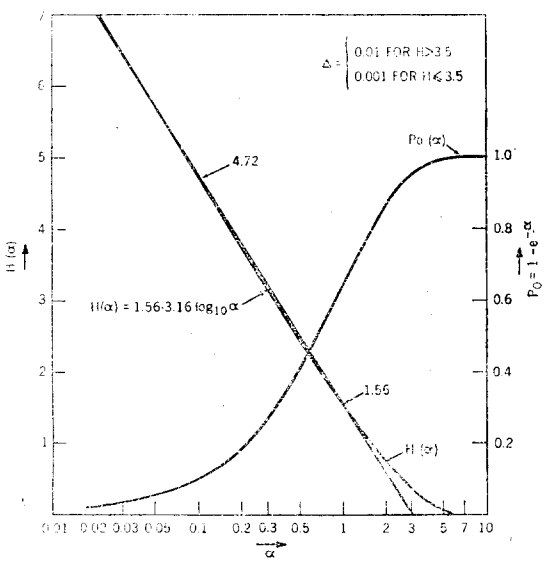

그림 2 Laplacian distribution parameter α에 對한 $H(\alpha)$, $H_a(\alpha)$ 및 $P_0(\alpha)$ 曲線.

Fig.2 $H(\alpha)$, $H_a(\alpha)$ and $P_0(\alpha)$ vs. Laplacian distribution parameter α.

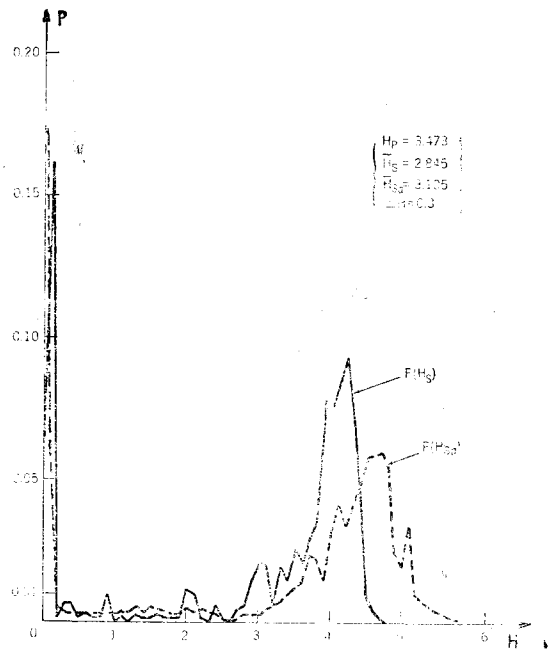

그림 3 測定한 線分 entropy H_s와 Laplacian 近似 線分 entropy H_{sa}의 分布曲線. 線分 길이 =256 samples.

Fig.3 probability densities for the measured segment entropy H_s and the Laplacian approximated segment entropy H_{sa} for segment length=256 samples.

않는다. $H_s<0.1$에서 높은 分布를 갖이고 있으나 이는 전체 平均 entropy에 큰 比重을 주지 못한다. 그림 3에서 H_{sa}는 대체로 H_s를 오른쪽으로 遷移한 모양을 갖이며 그 平均 差異값이 0.3(bit)라는 높은 값을 갖이는 것은, $\{d_n\}$이 Laplacian 分布를 갖이며, 또한 統計的으로 獨立이라는 假定이 不完全했기 때문이다. 그러나 이런 一貫된 變位값은 符號(code) 選別過程에서 보상할 수 있다. 사진 전체의 entropy H_p가 H_s의 평균값 $\overline{H}_s$보다 큰 것은 entropy의 convexity 성질에 의한 것이며, 이 差異값이 큰수록 adaptive coding이 큰 효과를 낼 수 있다.

4. 符號와 符號選別

4—1. 符 號

그림 3과 같은 entropy 分布를 갖이는 데이타에 여덟가지 符號(code)로서 높은 符號化 効率을 얻으려고 한다. 여기서 code set의 크기를 8로 한 것은 符號選擇 즉 適應性의 餘裕도 있고, 또 같은 길이인 3 bit로서 選擇된 符號의 識別이 편리하기 때문이다. 여덟가지의 符號를 構成하기 위해서는 우선 entropy를 여덟가지 區域으로 量子化해야 한다. 量子化를 위해서는 entropy 分布 曲線과 量子化基準이 확실해야 한다. 그러나 任意의 분포곡선에 대한 最適量子化는 쉬운 일이 아니다.[8][9]

여기서 限定되었기는 하나 주어진 자료인 그림 3으로 부터 주요 판단을 내릴 수 밖에 없다. entropy가 아주 작은 경우($H_s\leq 0.5$)는 사진의 검은 部分에 해당되는 것으로서 그 比重이 크므로 마땅히 한 符號(code)가 配定되어야 할 것이다. 그 以外에서는 최대값이 $H_s=4$ 근처인 unimodal curve를 나타낸다고 볼 수 있다. 여기서 한 段階 더욱 나아가 量子化의 효율과 위험율을 감안해서 Gaussian 分布에 대한 量子化 結果를 이용할 수 있겠다. 즉 最少 平均自乘誤差를 基準으로 한 optimum fixed quantization level(=7) quantizer를 생각할 수 있다.[8] entropy의 上限값이 $H_s=6$인 것을 참작하여 平均値 4, 標準偏差 0.9인 Gaussian 분포에 대한 optimum 7 level quantizer를 구하면 그림 4와 같다. 그리고 여기서 또 한번 便宜上 量子化 境界값(input quantization level)과 그 代表값(output quantization level)을 괄호안의 近似값으로 취한 것이 우리가 사용할 값이다. entropy가 제일 낮은 區域에서 그 代表값 $H_s=2.17$을 $H_s=1.5$로 크게 대신한 것은 그림 3에서 이 區域이 事實上 均一한 分布에 가깝기 때문이다. 以上으로 구한 최종적인 量子化 境界값은 다음과 같다.

$$H_s=0.5,\ 2.5,\ 3.25,\ 3.75,\ 4.25,\ 4.75,\ 5.5 \quad (12)$$

위의 境界값으로 區分된 各 entropy 區域에서는 그 區域 代表값에 대한 符號(code)를 공통적으로 사용해야 한다. 이 대표값 $H_s=0.3,\ 1.5,\ 3.0,\ 3.5,\ 4.0,\ 4.5,\ 5.0,\ 6.0$에 대한 符號(code)를 각각 $C_0,\ C_1,\ C_2,\ C_3,\ C_4,\ C_5,\ C_6,\ C_7$이라 두자. source coding에서는 Huffman 符號[10]가 효율이 가장 좋은 것으로 알려져 있다. 그러나 entropy가 1보다 작은 경우에는 그 효율이 매우 나쁘다. 따라서 符號 C_0는 Run Length Code[11]로 구성함이 適格일 것이나,[12] entropy가 아주 작은 이 경우는 사진 데이타에서 대개 검은 周邊地域에 해당하므로

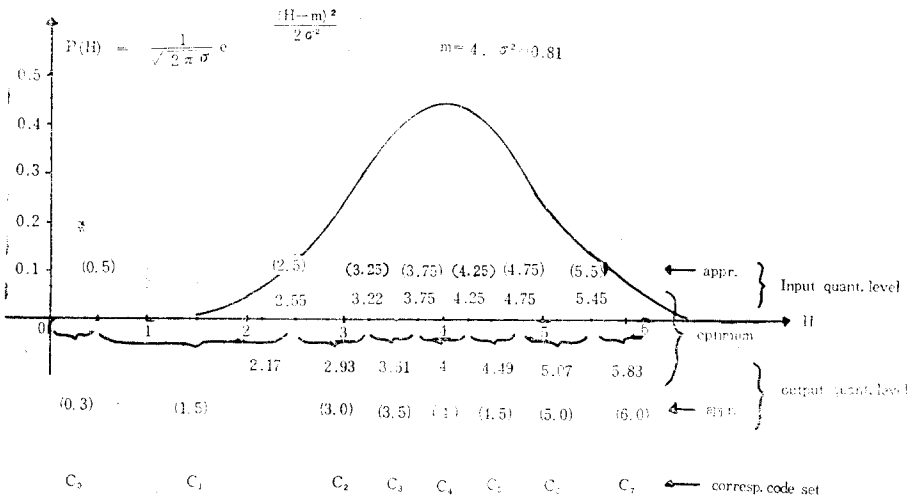

그림 4 entropy quantization 間隔
Fig.4 Entropy quantization intervals.

d	$C_1: H=1.5,\ p_0=0.65$	$C_2: H=3.0,\ p_0=0.292$	$C_3: H=3.5,\ p_0=0.212$	$C_4: H=4.0,\ p_0=0.154$
0	1	10	10	100
±1	01	11	11	101
±2	001	010	010	110
±3	0001	011	011	111
±4	00001	0010	0010	0100
±5	000001	0011	0011	0101
±6	0000001	00010	00010	0110
±7	00000001	00011	00011	0111
±8	000000001	000010	000010	00100
±9	0000000001	000011	000011	00101
±10	00000000001	0000010	00000100	00110
	00000000000 ×××××× −10	0000011	00000101	00111
±11		00000010	00000110	000100
±12		00000011	00000111	000101
±13		000000010	000000100	000110
±14		000000011	000000101	000111
$\lvert d\rvert \geq 15$		00000000100	000000110	0000100
		00000000101	000000111	0000101
		00000000110	0000000100	0000110
		00000000111	0000000101	0000111
		000000000 ×××××× −11	0000000110	000001000
			0000000111	000001001
			00000000 ×××××××× −14	000001010
				000001011
				000001100
				000001101
				000001110
				000001111
				000000 ××××××

表 1-1 補完된 Shannon-Fano 符號 $C_1 \sim C_4$

Table 1-1 Modified Shannon-Fano codes C_1 to C_4 (××××××represents 6 bit sample value)

데이타를 充實하게 전부 보내지 않드라도 별 損失이 없음을 알 수 있다. 따라서 C_6를 符號(code)識別 bit 이외에는 더이상 情報를 보내지 않는 符號(code)로 하였다. 反面에 $H_s=6$인 경우는 이 값이 바로 可能한 entropy의 최대값이므로 符號化를 더 할 必要없이 data bit를 그대로 傳送하는 것이 마땅하다. 즉 符號 C_7은 PCM 符號 그 자체이다.

符號 $C_1 \sim C_6$는 複號(decoding)의 便宜上 Huffman 符號보다 Shannon-Fano 符號[13]를 택하였다. 여기서는 더욱 複號의 便宜를 위해서 各 符號(codeword)中에서 같은 接頭 bit를 갖인 것들은 接尾 bit의 길이가 갖도록 補完한 Shanon Fano 符號(modified S-F code)를 사용하였다(表 1). 各 符號의 구성절차는 주어진 entropy에 대해서 (7), (8), (9)式에 의한 確率分布를 먼저 구한 다음, 위의 補完型 Shannon-Fano 符號를 구성한다. 이 때 같은 確率分布에 대한 Huffman 符號의 各 符號(codeword) 길이와 같은 符號(codeword) 길이를 갖도록 노력하면 Huffrnan 符號에 近似한 符號化效率을 얻을 수 있다. 個別的인 符號(codeword)를 配定하는 差異값의 上限 M은 (8)식으로 定해지나 같은 길이의 接尾 bit를 갖도록하는 補完過程에서 表 1에서와 같이 $(M\pm1)$로된 경우가 있다. 이것은 entropy 계산에는 別無 영향이지만 符號化, 複號化에서는 매우 편리한 점이 된다.

以上으로서 adaptive coding에 使用될 符號(code)는 두가지 等長符號(constant length code)와 非等 長符(variable length code)인 여섯가지 補完型 Shannon-Fano 符號로 構成되었다. adaptive coding과 比較하기 위해서 사진데이타 전체의 確率分布를 측정한 다음 이에 맞는 補完型 Shannon-Fano 符號를 構成한 것이 表 2 와 같다.

4—2 符號選別

주어진 code set 중에서 데이타의 변화하는 統計的 特性에 가장 適合한 符號를 찾는 것이 本 論文의 adaptive coding 方法에서는 매우 중요한 과정이다. 주어진 線分(scan line segment) 데이타에서 entropy를 측정한 다음, 前節에서 구한 entropy 量子化 境界값과 比較해서 適合한 符號를 찾는 것이 符號化 效率上 가장 좋은 방법이다. 그러나 entropy 대신에 $P_0=P[d_n=0]$을 符號選別의 기준으로 사용하면 훨신 신속한 符號 選別이 가능하다. 이는 $P[d_n]$이 Laplacian 分布에 近似하다는 前提를 活用한 것으로서 entropy 경우보다 符號化 效率은 低下된다. 그러나 real-time 符號化를 위한 所要時間 節約이라는 큰 장점이 있다.

	$C_5: H=4.5,\ p_0=0.11$	$C_6: H=5.0,\ p_0=0.0787$
0	100	1000
	101	1001
±1	110	1010
	111	1011
±2	0100	1100
	0101	1101
±3	0110	1110
	0111	1111
±4	00100	01000
	00101	01001
±5	00110	01010
	00111	01011
±6	000100	01100
	000101	01101
±7	000110	01110
	000111	01111
±8	00001000	001000
	00001001	001001
±9	00001010	001010
	00001011	001011
±10	00001100	001100
	00001101	001101
±11	00001110	001110
	00001111	001111
±12	000001000	0001000
	000001001	0001001
±13	000001010	0001010
	000001011	0001011
±14	000001100	0001100
	000001101	0001101
±15	000001110	0001110
	000001111	0001111
±16	0000001000	00001000
	0000001001	00001001
±17	0000001010	00001010
	0000001011	00001011
±18	0000001100	00001100
	0000001101	00001101
±19	0000001110	00001110
	0000001111	00001111
±20	0000000××××××	000001000
		000001001
±21		000001010
		000001011
±22		000001100
		000001101
±23		000001110
		000001111
±24		0000001000
		0000001001
±25		0000001010
		0000001011
±26		0000001100
		0000001101
±27		0000001110
		0000001111
±28	−23	0000000××××××
$\|d\|\geq 29$		

表 1-2 補完된 Shannon-Fano 符號 C_5, C_6
Table 1-2 Modified Shannon-Fano codes C_5, C_6 (×××××× represents 6 bit sample value)

그림 3은 實測한 entrory와 Laplacian 分布라는 近似値와의 差異를 잘 보이고 있다. 대체로 $H_s>2$인 구역에서는 같은 P_0값에 대해서도 Laplacian 分布라는 假定에서 얻은 entropy H_{sa}는 實測한 entropy H_s 보다 平均 0.3 bit가 높다. 따라서 P_0를 符號選別의 基準으로 사용하려면 0.3 bit만큼 增加된 Laplacian entropy에 대한 量子化 境界값을 만들 필요가 있다. 즉 (10), (7)식으로 부터

$$\alpha(H)=10^{(1.56-H)/3.15} \tag{13}$$
$$P_0(H)=1-e^{-\alpha(H)}$$

이 되며, 여기서 entropy H를 $(H+0.3)$으로 代置한 後에 (12)식의 各 entropy 量子化 境界값을 代入하면 이에 相應하는 P_0 量子化 境界값을 얻을 수 있다. 그림 3에서와 같이 entropy가 아주 낮은 때에는 H_s와 Laplacian entropy H_{sa} 사이에 別 差異가 없으므로, 제일 낮은 entropy 量子化 境界값 $H=0.5$에 對應하는 P_0값은 (13)식에서 그대로 얻을 수 있다. 以上으로서 實測 entropy와 Laplacian entropy 사이의 差異點을 補償한 P_0 量子化 境界값은 다음과 같다.

$$P_0=0.917,\ 0.333,\ 0.209,\ 0.150,\ 0.107,$$
$$0.076,\ 0.045 \tag{14}$$

d	codeword
0	1
1	0100
±2	0101
	0110
±3	0111
	00100
±4	00101
	00110
±5	00111
	000100
±6	000101
	000110
±7	000111
	0000100
±8	0000101
	0000110
$\|d\|\geq 9$	0000111
	00000××××××

表 2 A fixed code for ATS-3 picture data
Table 2 (×××××× represents 6 bit sample value)

5. 實驗結果

前節에서 求한 code sets와 符號 選別을 위한 entropy와 P_0의 量子化 境界값을 이용한 模擬 實驗結果는 그림 5, 그림 6과 같다.

그림 3에서와 같이 線分 entropy H_s는 線分길이 256 sample에 대한 測定値이다. 그림 3은 그림 1의 全 走査線에 대한 結果인 反面에 그림 5와 그림 6 에서는 컴퓨타 時間을 節約하기 위해서 每 10番째 走査線을 처음부터 차례로 120走査線에 대한 것만을 취한 결과이다. 따라서 그림 1의 사진에서 윗 半面만의 符號化 結果를 얻은 셈이다.

그림 5는 符號化로 얻은 데이타 減縮比(compression ratio)를 나타낸다. 符號化하기 前의 data는 每 sample

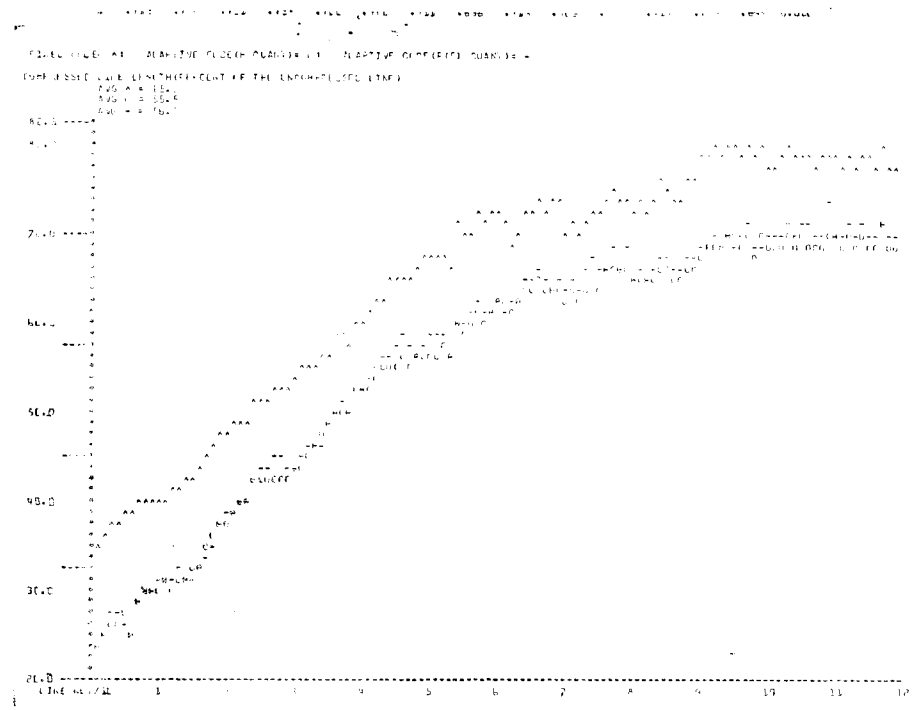

그림 5 走査線 別 平均 데이타 減縮比
Fig.5 Average data compression ratio vs. scan line.

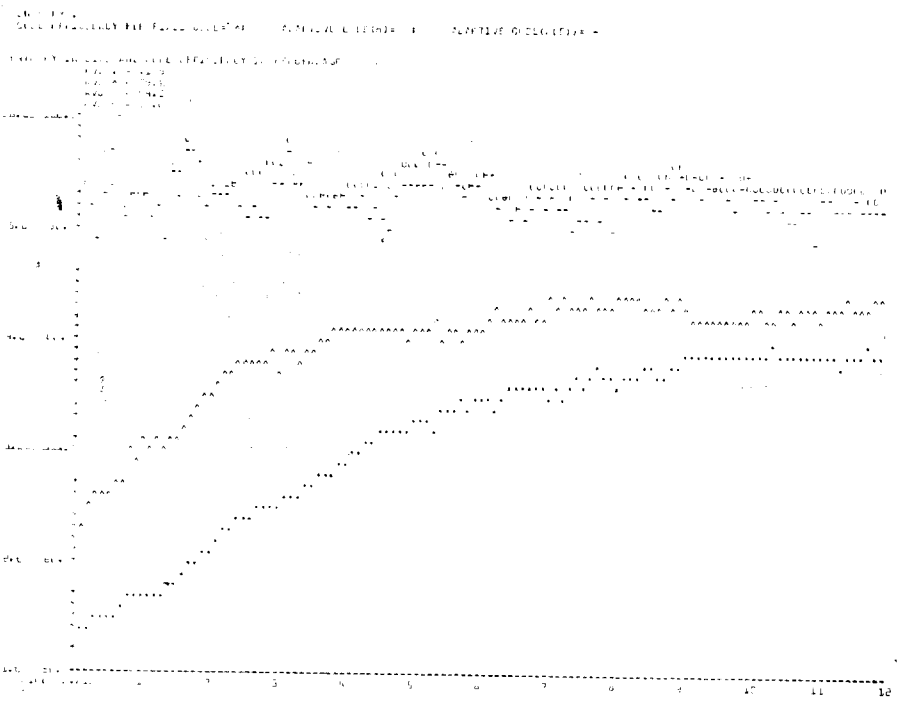

그림 6 走査線 別 平均線分 entropy H_s와 平均 符號化効率
Fig. 6 Average segment entropy H_s and average code efficiency vs. scan line.

마다 6bit로 表示되어 있었으므로 減縮比는 다음과 같다.

減縮比=平均 sample 길이/6 (15)

그림에는 entropy 量子化와 P_0 量子化로 符號選別한 adaptive coding과 表 2의 符號로 된 fixed coding때의 走査線 別 平均 減縮比를 圖示하였다. 그림에서 볼 수 있는바와 같이 fixed coding 때의 減縮比曲線은 대체로 adaptive coding 때의 減縮比曲線을 위로 10% 옮긴 曲線이다. entropy 量子化를 이용한 adaptive coding 때와의 平均減縮比 差異는 9.5%이다. 두 adaptive coding 사이에는 0.8%의 差異밖에 없다.

그림 5에서 모든 減縮比曲線은 走査線에 따라서 크게 變하고 있다. 즉 減縮比는 entropy값에 따라 變함을 알 수 있다. 이는 그림 6의 走査線 別 平均 entropy 曲線과 比較하면 쉽게 알 수 있다. 그림 6의 다른 세 曲線은 그림 5의 세 曲線에 對應하는 符號化效率을 表示하고 있다. 여기서 符號化效率(code efficiciency)은

符號化效率=entropy/平均 sample 길이 (16)

이다. fixed coding 때의 符號化效率 曲線은 그림 5의 세 曲線과 마찬가지로 entropy 曲線을 따라 變化하고 있으나, adaptive coding 때의 符號化效率 曲線은 entropy 曲線에 相關없이 그 平均 符號化效率 값에 近似한 모양을 하고 있다. 두 adaptive coding의 平均 符號化效率 사이에는 1.3%의 差異를 보이고 있다. adaptive coding에 의한 平均 符號化效率 改善은 各各 14.7%, 13.4%를 나타내고 있다.

그림 5와 그림 6의 제한된 測定結果에서도 變化하는 entropy에 無關한 parameter는 adaptive coding 의 平均 符號化效率임을 알았다. 따라서 이 平均效率과 그림 3의 全 走査線에 對한 平均線分 entropy $\overline{H_s}$로 부터 그림 1의 사진 전체에 대한 데이타 減縮比를 구할 수 있다. 그림 6에서 entropy 量子化와 P_0 量子化때의 平均 符號化效率은 各各 E_H=94.3%, E_P=93%이며, 그림 3에서 平均線分 entropy는 $\overline{H_s}$=2.845이다. 이를 (16), (15)식에 代入하면 다음과 같이 두 adaptive coding 때의 데이타 減縮比 C_H, C_P를 구할 수 있다.

$$C_H=50.3\% \quad (17)$$
$$C_P=51.0\%$$

6. 結 論

前節에서 구한 code set와 符號選別 方法으로서 사진 데이타 全體에 대해서 約 50% 즉 2對1의 데이타 減縮比를 얻을 수 있다. 이 adaptive coding이 fixed coding에 比해서 데이타 減縮比와 符號化效率에서 約 10%와 15%의 利得이 있음은 이미 위에서 言及된 바와 같다. adaptive coding을 위한 두 符號選別 parameter인 entropy와 P_0 사이에는 데이타 減縮比와 符號化效率에서 1% 內外의 差異밖에 없다. 따라서 P_0는 所要時間 節約이라는 長點뿐만 아니라 performance도 우수한 符號選別 parameter임을 알 수 있다. 앞으로 더욱 檢討되어야 할 것은 difference data sequence $\{d_n\}$의 確率分布를 (7)식 보다 더욱 實際에 近似하도록 얻는 일이다. 여기서 分布函數를 Laplacian 分布보다도 Gamma 分布[9]로 보는 것이 또 한 方法이 될 것이다. 또한 $\{d_n\}$을 統計的으로 獨立的인 data sequence로 보는 것 보다는 (6)식의 결과에 좀더 가까운 一次 Markov sequence로 본다면 좀더 좋은 符號化效率을 얻을 수 있을 것이다.

謝 意

著者는 本 硏究遂行을 積極的으로 後援한 美國 NASA/GSFC의 Mr. Richard L. Kutz, 컴퓨타 實驗을 도운 中央電子計算所의 조기종씨와 韓國科學院의 박상인, 유영욱, 최동재군에게 깊은 感謝의 뜻을 表합니다. 또한 論文構成과 많은 用語修正을 모아주신 未知의 審査委員에게 感謝드립니다.

附 錄

difference data sequence $\{d_n\}$의 相關係數(correlation coefficient) : (5)式으로 부터 difference data sequence $\{d_n\}$은 다음과 같다.

$$\{d_n\}=\{s_n-s_{n-1}\} \quad (A-1)$$

autocorrelation 함수와 normalized autocorrelation 함수를 各各 $R(i)$, $\rho(i)$로 表示하면 주어진 data sequence $\{s_n\}$과 difference data sequence $\{d_n\}$에 대한 값은 各各 다음과 같다.

$$R_s(i)=E[s_n s_{n-i}]$$
$$R_d(i)=E[d_n d_{n-i}] \quad (A-2)$$
$$\rho_s(i)=R_s(i)/R_s(0)$$
$$\rho_d(i)=R_d(i)/R_d(0) \quad (A-3)$$

여기서 E는 統計的 平均(statistical expectation)을 나타낸다. $\{s_n\}$이 stationary sequence이라 보면 $E[d_n]=0$ 이므로 (A-3)식의 normalized autocorrelation은 바로 $\{d_n\}$의 相關係數가 된다.

(A-3)式을 展開하면

$$\rho_d(i)=\{E[(s_n-s_{n-1})(s_{n-i}-s_{n-i-1})]\}/R_d(0)$$
$$=\{2R_s(i)-R_s(i-1)-R_s(i+1)\}/R_d(0) \quad (A-4)$$

여기서 또한

$$R_d(0)=E\{(s_n-s_{n-1})^2\}$$

$$=2\{E[s_n^2]-E[s_n s_{n-1}]\}$$
$$=2R_s(0)\{1-\rho_s(1)\} \quad (A-5)$$

이다. 따라서 (A—4), (A—5)식을 보으면,

$$\rho_d(i)=\{\rho_s(i)-[\rho_s(i-1)+\rho_s(i+1)]/2\}/[1-\rho_s(1)] \quad (A-6)$$

difference sequence를 (A—1)式과 같이 둔 이유는 $\{s_n\}$의 correlation이 매우 높다는 實際的인 假定때문이다. correlation이 높으면 correlation 함수의 變化率이 매우 낮아진다. 따라서 $\{\rho_s(i)\}$ sequence에서 이웃 i의 값에 대해서 $\rho_s(i)$는 것의 같은 값이다. 이런 性質을 (A—6)식에 適用하면 다음과 같은 結果를 얻는다.

$$\rho_d(i)\cong\begin{cases}-0.5, & i=\pm1\\ 0, & i=\pm2,\ \pm3,\ \cdots\cdots\end{cases} \quad (A-7)$$

參 考 文 獻

1. J. Ziv, "Coding of sources with unknown statistics," *IEEE Trans. on Information Theory*, vol. IT—18, no. 3, pp. 384—394, May, 1972.
2. L. D. Davisson, "Universal noiseless coding," *IEEE Trans. on Information Theory*, vol. IT—19, no. 6, pp. 783—795, Nov. 1973.
3. R. F. Rice and J. R. Plaunt, "Adaptive variable length coding for efficient compression of spacecraft television data," *IEEE Trans. on Communication Technology*, vol. COM—19, no.6, pp. 889—897, Dec. 1971.
4. L. D. Davisson and R. L. Kutz, "An operational data compression system using minicomputers," *Record of* 1972 *National Telecommmnications conference*, pp. 34A—1—34A—5, Dec. 1972.
5. R. G. Gallager, *Information Theory and Reliable Communication*, New York: Wiley, 1968, p. 57.
6. A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, New York: McGraw-Hill, 1965, p. 420.
7. J. B. O'Neal, Jr., "Predictive quantizing systems (DPCM) for the transmission of television signals", *The Bell System Technical Journal*, vol. 45, pp. 689—721, May/June, 1966.
8. J. Max. "Quantizing for minmun distortion," *IRE Trans. on Information Theory*, vol. IT—6, pp. 7—12, March. 1960.
9. M. D. Paez and T. H. Glisson "Minimum mean-squared-error in speech PCM and DPCM systems," *IEEE Trans. on Communication*, vol. COM—20, no. 2, pp. 225—230, Apr. 1972.
10. D. A. Huffman, "A method for the construction of minimum-redundancy codes," *Proc. of IRE*, pp. 1098—1101, Sept. 1952.
11. S. W. Golomb, "Run-length encodings," *IEEE Trans. on Information Theory*, vol. IT—12, no. 3, pp. 399—401, July, 1966.
12. Jae-kycon Kim and R. L. Kutz, "A study on adaptive data compression for the cloud pictures of ATS—1, ATS—3 and ITOS-D," X-document of NASA/GSFC, Greenbelt, Md. April, 1973.
13. R. M. Fano, *The Transmission of Information*, New York: 1961.