社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2008－CVIM－162 (82)
2008／3／11

エントロピー計測におけるサンプル計測の有効性について

大関和夫 芝浦工業大学

MPEG 圧縮データの乱数の度合いを調べ、再圧縮の可能性を検証している。長いビットの事象に対し、エントロピーが低下したり、相関が観測されたりすれば、その事象に対して再圧縮が可能となる。長いビットの事象を短いビットにサンプルした時、初めの長いビットの事象のエントロピーはどのように変化するかの対応がわかれば、サンプル計測の妥当性が示される。数ビットのいくつかのビットパターンからビット数を削減したサンプルを求め、それらのエントロピー値を比較し、短い方から長い方を推定する事後推定の有効性を検討し、実験の手法を提案する。

Validity of Sample Measurement in Entropy Measuring

Kazuo Ohzeki (Shibaura Institute of Technology)

Possibility of re-compression of MPEG coded data is investigated using random number analysis. To obtain N-gram Shannon entropy tends to be difficult as the value N increases. It is difficult to measure the entropy for events with longer bit elements, but it is easy to do so for events with shorter bit elements. It is important to clarify the probability correspondence between two observations for of the same event with elements as the longer bit elements and shortly sampled elements. Using the correspondence of probability, N-gram Shannon entropy with longer bit elements can be estimated from the results of N-gram Shannon entropy with shorter bit elements. Using a several patterns of bit events, correspondences with mean and variance and maximum width of entropy are calculated.

1. はじめに

MPEG などの画像符号化国際標準方式は，事実上最適な方式とみなされて来た．しかし，その理論的な裏付けは難しく，最適性が保証されているわけではない．動き補償，イントラ／インター，ＤＣＴ，ＶＬＣなどの個々の要素の性能向上が図られ，要素内ごとの最適化が目指されて来た．個々の最適化の残りや要素間の相互情報量などの集積については未検討で，全体的な最適化に関わる検討は R/D 最適化などでなされているが，ヒューリスティックな手法を行っていることには変わりはない．

ここでは，圧縮後のビットストリームを解析してその冗長度を評価することにより，圧縮アルゴリズムの完成度を外部から評価していくことをめざす．また扱い易い冗長性があれば、再圧縮を行うことが可能になる．この冗長度は，出力系列を長い区間観測した時，現れる可能性がある．圧縮ビットストリームを解析するためには，十分長い区切りから成るビットの事象を調べる必要があるが，状態数が指数的に増大していくため，20-30 ビットが現状では第一の限界である．したがって，長い区切りの事象を想定した上で，短い区切りでのサンプルにした事象の結果から長い区切りの事象の確率を推定するときの有効性を調べることを行う．

出力ビットストリームの解析から符号の系列の冗長度を評価することを提案し，各種計測が行われて来た[1-5]．これまで MPEG-2 を主体として符号化系列の冗長度解析を行うことによって，符号化性能の評価がなされて来た．従来，データの冗長度の検証として多くの乱数検定方式が試みられて来た．文献[6]には十種近く列挙されているが，それぞれ何らかの観点から選定された頻度等の一様性を調べることを行うものが多い．Gail Gasram は Diehard なる乱数検定ソフトを開発している

[7]．また，John Walker はインターネット上で系列を入力すればそのエントロピーを計算するプログラムを公開している[8]．しかしながら，計測対象は 1 バイトの単位であり，符号化ファイルなどの長区間のデータは対象としていない． H.Moradi らは英文テキストの数十文字という長期エントロピーを計測しているが[9]，単語や短文を固定してはじめからパターン数が削減されており，データ総量が少なく，すべてのパターンを尽くす解析を志向していない．J. Soto は乱数発生とテストについて実験的手法で比較検討を行っている[10]．これら乱数の検証は，またより完全な乱数の発生方式のためでもあり，パターンを一様化するための工夫が行われている．現実には計算上の制約から数ビットのパターンを調べるなどの方式が多い．符号化ビットストリームの解析においては長区間のビットパターンの解析を行い，その中に冗長度がどのように含まれているかを見出す必要がある．エントロピー計測はある程度まで力づくで行うことも可能だが，膨大な事象を全て調べることは不可能なので，分布の形状を解析し，モデル化できる部分は理論的モデルによる検証を導入する必要がある．

符号化ビットストリームのエントロピー，サンプルエントロピー等を調べ，長いビット数ごとに区切ったビットパターンという統計量では 1 ビットあたりのエントロピーが 20-30%低下し，再圧縮の可能性ある事が示されて．一方，この統計量はビット長が 1 ビット増加する毎にデータ量が 2 倍となり，対象としている圧縮ファイルのデータ量は指数関数的に多く確保する必要があった．データ量とともに符号化対象の映像や符号化パラメータの種類に関しても十分多くの数を用意した上でエントロピー計測を行うことにより，結果の精度が確保されることになるが，どの程度のファイル量を用意すれば普遍的で信頼性ある結果が得られるかは不明であった．本論文では，ビットストリーム解析の手法の検証を行い，計測結果の普遍的な信頼性を確保するために必要なビットストリームの長さ，種類の数を明確化していく．ビットストリームの長さについては，事象の種類の数倍が必要である事が分かった．また，種類に関しては，計測対象である確率事象をモデル化し，その分布の様態を調べた．ある頻度の集合に対する反転分布を考えると，その集合と反転分布の和集合はエントロピーが 1 となる．等エントロピー空間を形成する分布はほぼ均一に広がっている．一方，MPEG 圧縮ファイルのサンプルの分布はやや特定領域に片寄っていることが，サンプルの結合ファイルのエントロピーを計測することで示されている．このサンプルによる検証は事象の数に連動して多く行う必要が無く，少数のサンプルで十分な結果を得ることが出来る．これらの検証手法では全ておよそ 20 ビットまでの長さのエントロピー計測を行っているが，より長いビット系列の計測が必要となっている．

本報告では，より長いビットの計測結果を推定により求めるための対応を明らかにし，その計測の手法を提案していく．

2. ビットストリーム解析

ある画像が符号化され一つのビットストリームのファイルとなる．一般には画像全体の集合は符号化され，ビットストリームのファイル全体の集合となる．ビットストリームのファイル全体の集合を扱うにあたり，その観測方法を定義し，その観測方法による統計量を評価すれば，符号化性能のある種類の最適性が評価できる．

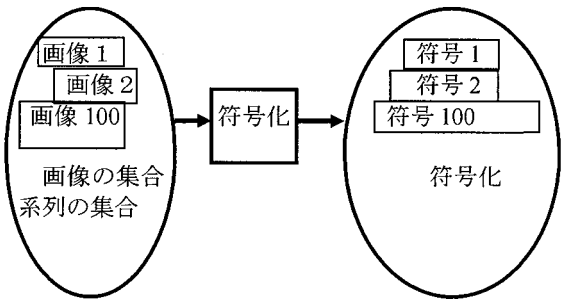


図 1 画像と符号化ビットストリームの集合
Fig.1 Sets of Images and coded Bit-streams

現実的な解析手段として，符号化ビットストリームを一定区間長のデータの集合とみなせば，現在，区間長が20ビット程度までなら容易に観測が可能な統計データとなり得る．この観測した統計データが不均一であれば冗長度が抽出できる．これは一種の乱数検定であり，定義した統計データの分布が均一であれば，その評価において乱数であり，不均一であれば乱数でない証拠を提示できることになる．符号化方式の構成は符号化要素の組み合わせから成り立っているが，出力ビットストリームの統計データの解析から，発生量の多い符号を生成している符号化要素が推定でき，これを符号化アルゴリズムの改善指針となる可能性がある．また，改善すべき符号化要素が複数の要素にまたがっている時は，各符号化要素の改良を行うことが難しいこともある．その場合，符号化ビットストリームの方を再符号化すれば，トータルで冗長度を効率的に削減できることになる．

符号化ビットストリームの解析には，固定区間長のデータのエントロピー計測[1] [2]，マルコフエントロピー計測[3]，条件付きエントロピー計測，サンプルエントロピー計測[3] [4]，自己相関計測[2]など多数の手法が検討されてきた．また，ストリームの計測では，従来の英文テキストのエントロピー計測[9]や，符号化データをバイト扱いとしたエントロピー計測[8]があったが，いずれも計測するデータ量は少なかった．

2.1 乱数の仮定

図2に示すように，理想的な符号化器で符号化された符号化ビットストリームは，再符号化が不可能となるべきものであるため，その集合全体の分布は均一で，集合の要素に対するエントロピーは1ビット当たり1となっているはずである．符号化ビットストリームの全体の集合は膨大なものであるので，ここでは取り扱いが可能な規模においても，この均一性の仮定が成り立っているものとする．

すなわち，再符号化の可否の観点から，符号化ビットストリームをある長さのビットパターンに区切り，そのビットパターンの集合は，その要素が均一の発生確率を有するものであるという仮定を設定する．この集合の全要素を1回ずつ取り出し一列に並べたとすると，その系列は長さが5の場合であれば，ポーカー検定[6]となり，これを拡張した区間検定法による乱数の理想形に一致しており，その意味で乱数をなす集合と呼ぶことにする．

このような考え方に基づき，符号化出力を可逆に再符号化するための条件として，以下のような乱数仮定の定式化を行う．再符号化不可能という意味で理想的な符号化器の出力をある長さのビットに区切ったビットパターンの発生確率は均一で，エントロピーは1ビット当たり1である．逆に不均一であればエントロピーが1ビット当たり1未満で再符号化が可能となる．1ビットあたりのエントロピーが1であるビットパターンの集合は，その要素を一列に並べると乱数となる．

本論文では以下，このように考え，その分布が均一のとき，ある区間長で区切ったビットパターンの集合を，簡略的に区分検定乱数と呼ぶことにする．また，検証を行おうとしている対象の符号化器を上記理想的と仮定し，その仮定を用いてエントロピーが1未満となるような集合を導き，背理法によって，再符号化が可能であることを示していく．

2.2 エントロピー計測

多数の乱数検定手法の中から，数値化が明確で，ビット長に柔軟性のあるエントロピーを指標として用いることを主とした．

計算機でエントロピーを計測していくと現状では，20ビット程度を超えると，メモリ容量の超過や演算時間の長期化という問題が発生する．

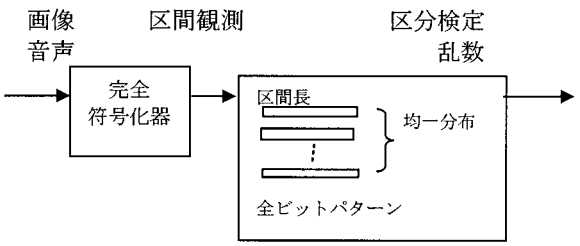


図2　完全符号化器の入出力
Fig.2 Input and Output of Ideal Encoder

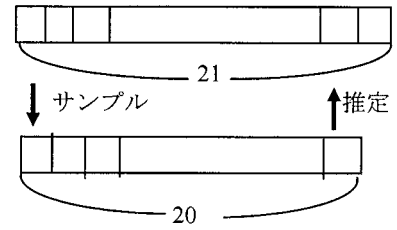
2.3 T-エントロピー

T-エントロピーは Titchner[11]により提案された T-Complexity を基に求めることが出来る指標で，動画像や MPEG-2 圧縮ファイルに対する計測がなされている[12]．図 3 は T-エントロピーの説明図で，0,1 の 2 分岐から出発し，過去に出来上がった符号化を接頭語として，系列を逐次増加していく．計測方式で、系列を逐次分解し拡張する T-augmentation により符号を作成していく。

3. サンプルエントロピーの計測

図 4 に今回検討する，長いビットのデータとそれを短くサンプルする時のデータの例を示す．上の例は 21 ビットのデータに対し，端の 1 ビットを削減し，計測する場合で，単純に 20 ビットのデータに対し、1 ビットの拡張の効果を評価するためのものである．下の例は，40 ビットのデータに対し，中の一部を除き，合計の長さが 20 ビットになるようなサンプルをとる場合で，40 ビットにわたる長い関係がある時には，確率分布に特に変化が発生し，特定の長周期を表出することができることにつながる．

このようにビット長に対するサンプル計測の手法は，もとの長いパターンに特殊な確率分布があれば，サンプルしたデータにもその形跡が残っていると考えることによる．サンプルすることによって，全くランダムなパタ



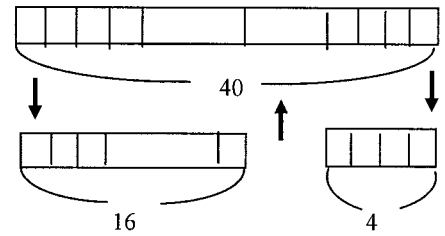


図 4．長ビットデータとサンプルの関係

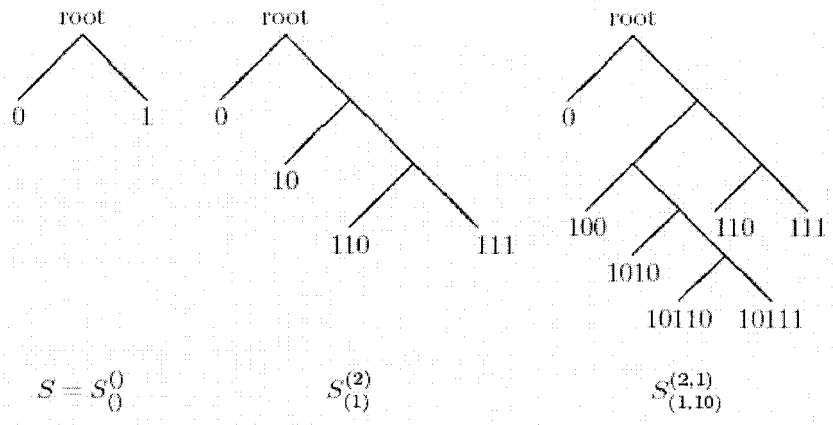


図 3． 逐次コピー増加の復号木を形成する T-augmentation[13]

ーンしか残らないケースはどの程度あるのか？を調べることにより，サンプル計測の有効性が求まる．また，ある確率的な対応関係が決定できれば，そのような状況の下にサンプル計測値を事後確率として，原因となる長いビット系列の確率密度を推定することの妥当性が示される．

長い系列とサンプル系列の組み合わせの関係は無限に近くあると考えられるので，ここでは以下のような単純な場合を設定し，その全パターンについての確率変化をエントロピーの計測値の変化として求めた．

3.1 実験

実験では，数ビットのパターン（長いビットのデータ）から成る集合を用意し，そのパターンから1ビット少ないパターンをサンプルパターンとして取り，そのエントロピーの変化を観測した．はじめのパターンは完全な乱数（つまり，3ビットなら000から111までの8パターンが均等に発生するもの，4ビットなら16パターンが均等に発生するもの）が出発点だが，計測の対象となるMPEG-2圧縮ファイルは完全な乱数ではなく，またその非乱数性を計測することを目標としているので，完全な乱数から任意の個数のパターンを除いたパターンについても合わせて実験を行った．この時，パターンのとり方は，たとえば3ビットの場合で，上記000から111までの8パターンやそこから削減されたn<8パターンが均等でない発生確率で存在する場合が一般的だが，ここでは各n<8に対し，均等に発生する場合のみをはじめの（長い）ビットパターンとして使用した．

例えば，3ビット8パターンのうち，110のみを除いた7パターンをはじめのデータとする時，先頭から2ビット取り出すサンプルは，次のようになり，その出現確率は2/7と1/7が混在している．しかし，後ろから2ビット取り出すサンプルでは，確率分布の配分が異なる事象がえられる．

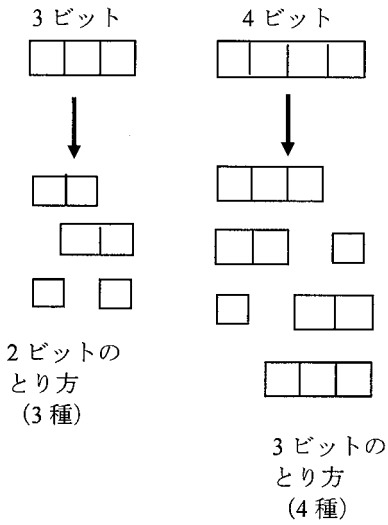
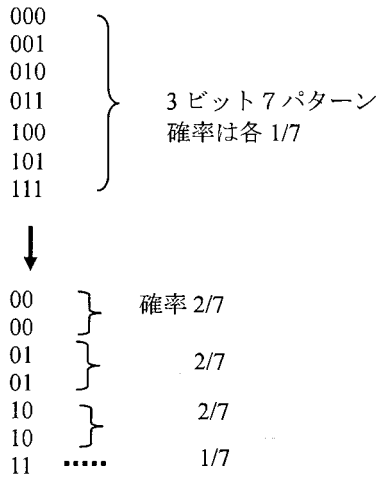


図5． 3ビットパターンから2ビットのとり方3種



このようなやり方で，n=3,4について行った結果が，図6，図7である．図6でパターンが8種の時は、全パターンがある場合で、前記完全な乱数の場合である．従って，2ビットのサンプルについても全パターンが発生し，推定できる特徴を有しない．n<8では，はじめのパタンのエントロピーも3未満になっているが，2ビットにサンプルされたパターン

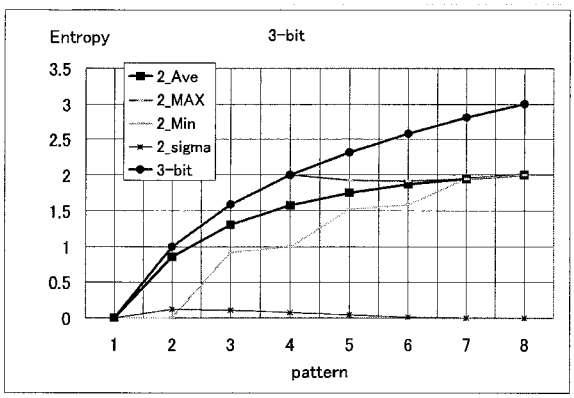


図6．3ビットから2ビットのサンプル生成

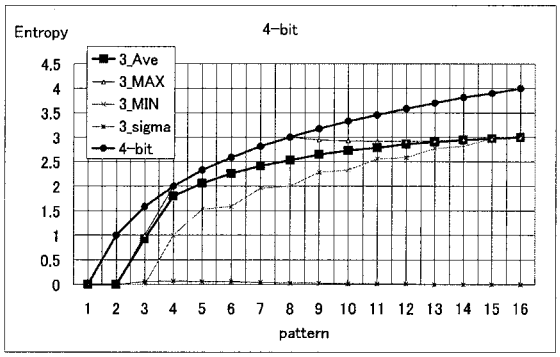


図7．　4ビットから3ビットのサンプル生成

も平均(Ave)で2ビット未満になり，n<7では最大値と最小値の間には幅がある．これから，少なくとも，大局的に，計測エントロピーが2未満になった場合は，元の3ビットのパターンのエントロピーも3未満になることがわかる．また，4ビットについてもより細かい特性が得られており，はじめのパターンのエントロピーが乱数でなく，その95%以下なら，サンプルデータのエントロピーも低下していることが確認できる．これから，簡単に例えば3ビットエントロピーを計測して平均2.9である時，はじめの4ビットパターンのエントロピーは3.7であったと推定することが可能となる．

4．　考察とまとめ

ビットパターンとより短いサンプルパターンのエントロピーを対応して計測しておくことにより，はじめのビットパターンに相関関係がある場合に，サンプルしたパターンにも関係をつけることができ，サンプル計測結果から，はじめの長い系列のエントロピーを推定可能となる．全体傾向として，完全な乱数と称するパターンの場合は，サンプルデータも乱数となり，非乱数のパターンは，サンプルもエントロピーが低下する．低下の割合は総じてビット数の低下の割合より小さい．つまり，サンプルされたデータの方が，1ビット当たりに換算したエントロピーが大きくなる．

サンプルされたデータのエントロピーの平均値と分散が既知となるので，ある量のサンプルされたデータの計測結果が得られた時，サンプル前のデータのエントロピーを確率的に推定することが可能となる．

今後は，計測するビット長を20ビット程度まで長くすること，パターンの発生を各一通りでなく変動がある場合などについて調べ，データが安定しているかの検証を行っていく必要がある．

参考文献

[1]島村定春，遠藤寛和，大関和夫“圧縮ビットストリームの解析による符号化性能限界の評価”電子情報通信学会総合大会D-11-163，　2000.3
[2]島村定春，大関和夫“圧縮ビットストリームの解析におけるサンプル推定計算方式”電子情報通信学会ソサエティ大会D-11-23，2000.10
[3]島村定春，大関和夫“圧縮ビットストリームの解析におけるサンプル推定計算方式と評価”PCSJ2000.P-P1.12，2000，11.
[4]K.Ohzeki，S.Shimamura，”Analysis of Codec Bitstream To Find Out Satistical Correlations for Longer Intervel”Proc.WOSPA2000，1-16.Dec.2000
[5]K.Ohzeki，S.Shimamura，”Analysis of Codec Bitstream of Picture-Effects of Calculating Accuracy”PCS-2001，FP-1-18，April.2001.
[6] 伏見正則「乱数（up応用数学選書12)」.東大出版会.1989.
[7] http://stat.fsu.edu/~geo/diehard.html
[8] http://www.fourmilab.ch/random/
[9]H.Moradi et al.，”Entropy of English Text: Experiments with Humans and a Machine Learning System Based on Rough Sets”，Information Sciences an international vol. 104，no.1-2，pp.31-48. 1998.
[10]Juan Soto, Jr.,, “Statistical Testing of Random Number Generators”, 22nd National Information Systems Security Conference, Applications papers: three
[11]M.R.Titchener, “Deterministic computation of string complexity, information and entropy”, Inter. Symp. On Inform. Theory, Aug. 16-21, 1998, Boston.
[12]K.Kawaharada, K.Ohzeki and U. Speidel, “Information and Entropy Measurements on Video Sequences”, Proc. of ICICS2005 pp.1150-1154, 2005.
[13]Ulrich Speidel, “T-Complexity and T-Informati on Theory”, CDMTCD Research Rept. Univ. of Auckland, Oct 2006.