解説

画像圧縮：画像内容の理解による符号化効率の向上

Image Compression: Improvement of Coding Efficiency by Understanding Image Contents

栄 藤 稔* *松下電器産業（株）中央研究所

Minoru Etoh* *Central Research Laboratories, Matsushita Electric Industrial Co., Ltd.

1. はじめに

画像符号化技術は、画像を少ないデータ量で伝送あるいは蓄積する技術として発展してきた。図1に音声・動画の圧縮率を示す。図中で、符号化標準MPEG[1][†]は、30 [Mbps]から1.2 [Gbps]の動画を1/20から1/40に圧縮して伝送することを可能にしている。この解説では、MPEGを題材に動画画像符号化方式の規格化の流れと画像の内容を理解することにより符号化効率を向上させる試みを紹介する。そして、動画画像の発生過程、状況、タスクに応じた情報量の割り当て、情報源モデルの選択が画像理解に基づく画像圧縮として重要であることを述べる。

2. 動画画像符号化の基本的構成

現在の動画画像符号化の基本構成にはマクロブロック動き補償（MC）+離散コサイン変換（DCT）符号化と呼ばれる方式が採用されている（MC-DCT符号化と略す）。これを図2に示す。MC-DCT符号化では時間的に前後するフレームの符号化・復号化結果を用いて現フレームの予測画像を生成する。これを動き補償（Motion Compensation）と呼ぶ。このために、マクロブロックと呼ばれる縦横16画素からなるブロックについて最も誤差絶対値和が最小となる移動量を動きベクトルとして検出し伝送する。予測画像はマクロブロックを検出された動きベクトルに従って平行移動させることにより生成される（マクロブロック動き補償）。これにより、フレーム間の相関を利用した符号化が行われる。図2の例では、画面上平行移動している車の現フレーム画像はマクロブロック動き補償により生成されるため、新たに大きな符号量は発生しない。

一方、現フレームから画面に現れた場合は動き補償の効果

が得られない。したがって、新たに予測誤差（残差）として符号化する必要がある。残差には空間的冗長度が残っているため離散コサイン変換（DCT）により空間的冗長度を除く。そしてDCT係数を量子化することにより、視覚的劣化がない範囲で情報を落とす。この時点でデータは不可逆となる。最後に量子化されたDCT係数をハフマン符号化などエントロピー符号化により統計的冗長度を除く。なお、予測画像は符号化側と復号化側で一致する必要があるために、符号化側では復号化側で行われる逆量子化、逆DCTを経て復号画像を生成し、次の予測に備える。

このように動画画像符号化は時間的に連続するフレーム間

図1 メディアと圧縮率

図2 マクロブロック動き補償+DCT符号化

原稿受付 1998年5月7日

キーワード：Computer Vision, Image Coding, Source Model, MPEG, Motion Model

*〒619-0237 京都府相楽郡精華町光台3-4

*Souraku-gun, Kyoto

[†]本来は国際標準化機構（ISO）のなかで動画と音声の符号化方式を定めるグループ（Moving Picture Expert Group.）の略称である。転じて、MPEGが定めた符号化方式そのものを意味する。

の相関を利用する動き補償と、予測が当たらない残差の空間的・統計的冗長性を除く波形符号化[†]の組み合わせからなっている。

2.1 動画像符号化の情報源モデル

図2に示した動画像符号化をより一般化して捉えると図3になる。動画像符号化は、時間軸方向の統計的冗長性を除き予測画像を生成する分析合成部と、予測画像と入力画像との差分である残差画像を符号化する残差波形符号化部から構成される。分析合成部は入力画像領域 R (位置ベクトルの集合) 中の任意形状・サイズの部分領域 $\omega_i \subset R$ から、 $\sum_{(x,y) \in \omega_i} |I(x,y) - I'(x',y')|$ が最小となる参照画像への幾何変換 $T: (x,y) \rightarrow (x',y')$ ($I(\cdot)$, $I'(\cdot)$ は入力画像, 参照画像の輝度) を求め、幾何変換とそれが適用できる領域の記述から入力画像を合成する (図4参照)。動画像汎用符号化では、前フレームに対する分析情報だけで、すなわち幾何変換による予測だけで人の初期視覚で知覚されないレベルまで投影像を再構成することは困難であり、残差符号化とのハイブリット構成とするのが一般的である。残差波形符号化部には DCT を典型とする周波数変換やベクトル量子化が用いられる。

以上の分析合成部と残差波形符号化部を設計する際に、二つのモデルが必要となる [3]。一つは、動画像に対する人の初期視覚モデルである。人の初期視覚は例えば、エッジ近傍では輝度振幅に、色では青に、解像度では斜めに感度が低い特性を持つ。これをモデル化することにより残差符号化において画像再構成の精度を落とす。入力画像に対する信号再生の忠実度は落ちるが、主観的画質を損なうことなく符号量を大きく削減することが可能になる。

もう一つのモデルは、動画像をどのような情報の発生過程と捉えるかを与える情報源モデルである。表1に画像符号化の情報源モデルを示す [4]。MPEG では動き補償で ω_i を縦横 16×16 画素間隔の固定形状のブロックとしておのおの $1/2$ 画素精度の平行移動の幾何変換を行っている。これは、レベル3の情報源モデルを採用していることになる。動画像が平行移動するブロックから成り立っているとの仮定は実世界の近似でしかありえない。そこで、近年の符号化研究では投影像の振舞いをより忠実にモデル化した高能率符号化の実現を目標として来た。

レベル4~5の情報源モデルを用いた符号化は分析合成符号化として知られる [5][6]。この符号化では図2の分析合成部において、 ω_i を任意形状領域とし、さらに滑らかな動きベクトル場モデルや2~8パラメータで記述されるパラメトリック動きモデル (アフィン動きや平面動きモデル) に基づく幾何変換が行われる。領域と幾何変換が求まれば、入力画像中の部分領域 ω_i は参照画像中の対応部分

図3 動画像符号化の基本構成

図4 動き補償

領域からマクロブロック平行移動に比べてより正確に予測することができる。しかし、一方で符号化データは、

- 分析パラメータ：規則性（一様性）をもった動きを持つ領域の記述符号（例：輪郭チェーン符号や4分木符号）+ 領域の動きパラメータ
- 残差符号

の総和となることから、分析パラメータのオーバーヘッドを考慮する必要が生じてくる。動き補償を精緻化しようとするすると残差符号量は減るが分析パラメータの符号量が増える。動画像高能率符号化の成否は、ある情報源モデルに従って動きと形状に規則性をもつ動き領域が動画像中に存在するかどうか、そして存在する場合それを確実に抽出できるかどうか依存している。この規則性検出は、現在もコンピュータビジョンの重要課題であり、レベル4以上の情報源モデルを用いた方式は汎用符号化方式としての地位を得るには至っていない。現状では、平行移動するブロックによる近似が、汎用動画像符号化の情報源モデルとして利用されている。

2.2 MPEG4 による物体単位の符号化

従来の動画像符号化技術で実現されなかった画像内容（画像を構成する各物体領域）に対応したスケーラビリティ^{††}、編集、特殊再生の機能やコンピュータグラフィックスを含む合成画像符号化が符号化標準化活動 MPEG4 において規格化されようとしている [7]。MPEG4 では、図5に示すように物体領域とその前後関係を保持した階層表現が採用されている [8]。これは VOP (Visual Object Plane) 符号化と呼ばれ、物体は前後するプレーンに分解されて符号化される。VOP 符号化は

[†]実際の DCT, エントロピー符号化の平易な解説として文献 [2] がある。

^{††}符号化速度に応じて再生品質を可変とする機能

表 1 情報源モデルと符号化技術

レベル	情報源モデル	符号化情報	符号化技術
1	画素	画素輝度値	PCM
2	統計的に冗長な画素	画素相関輝度値	予測符号化, 変換符号化
3	平行移動するブロック	ブロック内相関輝度値, 動きベクトル	動き補償 + DCT 符号化
4	動きが異なる動領域	マッピングパラメータ 形状, テクスチャ	領域/テクスチャ符号化
5	構造が制約された物体領域	パラメトリック動き 形状, テクスチャ	分析合成符号化
6	3次元構造が既知の動物体	形状, 動き, テクスチャ	モデルベース符号化

図 5 階層画像 (VOP) 符号化

- 形状：フレーム間 2 値算術符号化 + 透過度多値表現
- テクスチャ：任意形状に対応した MC-DCT 符号化で構成され [9], 動画像を物体ごとに符号化することが可能となっている。

形状符号化とテクスチャ符号化にレベル 3 の情報源モデルが使われる一方で、スプライト (SPRITE) 符号化と呼ばれる画像合成機能が実現されている。スプライトはレベル 5 の情報源モデルに基づいて、プレーンを変形する処理であり、スプライト画像を一度伝送しておくで各フレームの再構成が 2~8 の幾何変換パラメータの伝送だけで可能となる。ここで重要なことは、シーンが前景と背景に構造化されて表現されることである。

図 5 の階層画像符号化 (VOP 符号化) は、物体領域の構造を保存して物体ごとに符号化を行う機能を持ち、さらにレベル 5 の符号化方式であるスプライトを包含している。これは画像符号化を単なる情報圧縮だけでなくシーンの持つ三次元構造を反映した符号化として捉えることができる。図 6 に、スプライトによる動画像伝送の例を示す。この例では入力画像の前景 (非剛体) はレベル 3 の情報源モデルを用いた VOP 符号化 (図中左下) として、背景 (剛体) はレベル 5 の情報源モデルを用いたスプライト符号化 (図中上) として実現され、復号時に前景と背景が合成される (図中右下)。なおこの例では、通常であれば 1.5 [Mbps] で実現される画質をスプライト符号化を用いて個別に符号化することにより 400 [Kbps] で実現している。ただし、スプライト符号化は このように “作られた絵” に対して有効であり、前述の分析合成符号化で指摘した領域抽出自動化の課題を解決するものではない。

図 6 スプライト例

MPEG4 は符号化標準として、構造化された物体単位の高効率符号化の枠組みを与えた点に意義がある。グラフィックスのオーバーレイ、CG と実写の融合に利用できる一方で、一般動画像符号化に MPEG4 の機能を利用するには、現時点で、オーサリングツールの利用が前提となる。

3. 画像理解と画像圧縮

コンピュータビジョンには、タスク実行のための選択的なシーン記述獲得が重要であるとする能動視覚の研究がある [10]。Ballard は Yabus の注視に関する文献 [11] を紹介して、シーン記述がタスクに依存することを示した [10]。そこでは、室内に複数の人間が登場する一枚の絵を被験者に提示して、「登場人物の年齢を述べなさい」という指示には、被験者の視線は絵中の人物の顔に注視され、一方「室内の登場人物の位置と家具を記憶しなさい」という指示には絵中のすべてをくまなく注視点が移動することが紹介されている。これは、視聴者（観察者）の意図/タスクにより注視する情報が異なり、保存すべき情報は個人とその状況により千差万別であることを示唆している。

一方、画像圧縮では図 3 に示したように、人の初期視覚で劣化が検知されない範囲で再生画像の忠実度を落している[†]。言い替えれば、初期視覚検知限内の情報はすべて

[†]画像圧縮では人の初期視覚検知限外の情報を落すことから、ボケ画像から距離を計測 (shape from defocus) するような人の視覚能力を越えた領域で圧縮画像を利用する際は注意が必要である。

図7 領域の重要度に応じた符号化

伝送することになる。以上の画像理解の視点で画像圧縮を眺めると以下の研究課題が見えてくる。

3.1 符号量の選択的割り当て

図7に符号量は同じだが、領域に応じて符号化精度を変化させた符号化例を示す。図中、左画像では均一な符号化精度とした結果であり、右画像では顔領域の符号化精度を上げ、逆に周囲を下げている。もし、視聴者（あるいはロボット）のタスクが人物の表情認識であり注視点が顔領域であれば主観的画質は右画像が良いことになる。そして同じ主観画質であれば右画像を左画像より少ない符号量で符号化することができる。

このように得たい情報、送りたい情報を、観察者あるいは製作者の意図に応じて選択的に伝送する画像メディアの形態が考えられる。応用を制約すれば、この取捨選択の操作を自動化できる。テレビ電話/会議では、通常、注視対象は顔であることから、顔領域自動追尾や色解析により、顔画像に多くの符号量を割り当てる符号化技術[12]も提案されている。この情報取捨選択の自動化はタスクと状況に依存した画像理解の問題である。

3.2 情報源モデルの自動選択

情報源モデルの選択は、分析合成符号化による高能率符号化に貢献するだけでなく、得たい情報、送りたい情報を選択する上で非常に重要な意味を持つ。

例えば、オーディオ符号化では音声と音楽の符号化アルゴリズムがまったく異なる。音楽符号化では音楽信号と聴覚特性の冗長性に注目して周波数変換に基づく情報源モデルを仮定するのに対し、音声符号化では、情報源が人の発声として特定できることから、声帯振動と声道の共振特性でモデル化する。入力信号が音声であることが分かり、背景雑音[†]を伝送しなくてよいのであれば、音声符号化を用いて10[Kbps]以下で品質良好な符号化が行える。しかし、そうでなければ符号化品質は破綻する。

MPEGに代表される動画像汎用符号化とモデルベース符号化[13]の関係はこの音楽と音声の関係に似ている。入力画像のある領域が顔であると分かれば、表1中レベル6の顔モデルを利用した高能率符号化が期待できる。タスク

と入力により符号化対象がなんであるかを識別して最適な情報源モデルが選択されるべきである。これも圧縮の問題ではなく画像理解の問題である。

4. む す び

2章、3章では画像符号化の基本構成とMEPG4を紹介した。レベル4以上の情報源モデルの利用は“汎用の高能率符号化”に貢献しているとは言い難い。一方、MPEG4は、オーサリングツールの使用を前提として画像をシーンの持つ三次元構造を反映した表現で記述することにより、符号化効率だけでなく、画像内容に応じた合成・編集などの新機能を実現する枠組みを提供している。

4章では画像圧縮にとって画像理解が重要であることを指摘し、今後の課題を述べた。汎用動画像符号化の範疇では圧縮率は、図1に示すように1/20から1/50の範囲にあり限界が見える。これはタスクに依存しないよう初期視覚検知限内の情報はすべて伝送する必要があるからである。これを越えるにはタスクに依存した動画像伝送の枠組みを考える必要がある。従来の画像符号化方式そのものを道具として使い、どのような状況で、どのような符号化方式、どのようなパラメータで適用すべきかを決定する制御の問題が今後重要となってくる。

ここに画像理解の役割が存在する。タスク実行のための符号量の自動/半自動割り当てなど制御技術への展開が期待される。また入力画像に対して情報源モデルを識別する問題がより大きな問題として存在する。

以上、画像圧縮が直面している課題は、コンピュータビジョンの課題と同根であることを理解していただければ幸いである。

参 考 文 献

- [1] D. Le Gall: "MPEG: A video compression standard for multi-media applications," Trans. ACM, 1991.
- [2] 安田浩, 渡辺裕: デジタル画像圧縮の基礎, 日経BP出版, 1996.
- [3] 原島 博: 画像情報圧縮, オーム社, 1991.
- [4] R. Schaphorst: "Draft report of the rapporteur for very low bitrate visual telephony," Technical report, ITU-TSS, SG15, WP15/1, Geneva, 1993.
- [5] H.G. Musmann, M. Hötter and J. Ostermann: "Object-oriented analysis-synthesis coding of moving images," Signal Processing: Image Communication, vol.1, no.2, pp.117-138, 1989.
- [6] P. Salembier, L. Torres, F. Meyer and C. Gu: "Region-based video coding using mathematical morphology," Proceedings of the IEEE, vol.83, no.6, pp.843-857, 1995.
- [7] "MPEG4 特集", 映像情報メディア学会誌, vol.51, no.12, 1997.
- [8] "Special issue on MPEG-4," IEEE Trans. on Circuits and Systems for Video Technology, vol.7, no.1, 1997.
- [9] "MPEG-4 visual committee draft 14496-3," ISO/IEC JTC1/SC29/WG11 Document, 1997.
- [10] D. Ballard: "Reference frames for animate vision," Proc. Int. Joint Conference on Artificial Intelligence, pp.1635-1641, 1989.

[†]雑音であるか否かは、状況による。話者の周囲状況を理解しようとすれば、背景音は雑音でなくなる。

- [11] A. Yarbus: "Eye Movements and Vision," Plenum, New York, 1968.
- [12] A. Eleftheriadis and A. Jacquin: "Automatic face location detection for model-assisted rate control in H.261-compatible coding of video," Signal Processing:Image Communication, vol.7, pp.231-248, 1995.
- [13] H. Harashima, K. Aizawa and T. Saito: "Model-based analysis synthesis coding of videotelephone images — conception and basic study of intelligent image coding—," The Transactions of the IEICE, vol.E-72, no.5, pp.452-459, 1989.

栄藤 稔 (Minoru Etoh)

1960年11月22日生。1983年広島大・工・第二類(電気系)卒業。1985年同大大学院博士課程前期修了。同年松下電器産業(株)に入社。1987~90年ATR出向, 1991~92年阪大受託研究員, この間コンピュータビジョンの研究に従事。その後MPEGに参加。現在, 松下電器 中央研究所主席研究員(チームリーダー)および奈良先端大客員助教授。メディア処理とネットワークの研究を指向。1993年信学会論文賞受賞。工博(阪大)。電子情報通信学会, 情報処理学会, IEEE各会員。