JPEG画像圧縮技術を用いた自然感の高い擬人化エージェント像の実時間生成

土肥　浩[†]　　二階堂信夫[†]
石塚　満[†]（正員）

Realtime Synthesis of Realistic Anthropomorphic Agent Moving Image Using JPEG Compression Technique

Hiroshi DOHI[†], Nobuo NIKAIDO[†], *Nonmembers*, and Mitsuru ISHIZUKA[†], *Member*

[†] 東京大学工学部電子情報工学科，東京都
Dept. of Information & Communication Engineering, Faculty of Engineering, The University of Tokyo, Tokyo, 113 Japan

あらまし　本研究では，さまざまな顔の向き，瞬き，口の形を変えた複数の顔画像をあらかじめJPEG形式に圧縮しておき，これを応答文やユーザの位置に合わせて実行時に任意の順序でつなぎ合わせることにより，高価なビデオアクセラレータを用いることなく，自然感の高い擬人化エージェント動画像の実時間生成を可能にした．

キーワード　擬人化エージェント，ユーザインタフェース，JPEG，インタラクション，顔

1. まえがき

我々は，ユーザとコンピュータとの自然なインタラクションを実現するため，自然感の高い顔画像と音声対話インタフェースを統合したビジュアルソフトウェアエージェント（VSA）[1]～[3]の研究を進めている．自然感の高い顔画像は，音声対話インタフェースにおいてコンピュータという“箱”に向かって話しかけるというユーザの心理的抵抗感を和らげる効果が期待できる．

自然感の高い顔画像を生成するためには，我々のVSAも含めて，テクスチャマッピング手法が用いられることが多い．しかし，この処理は計算コストが高く，ヒューマンインタフェースとして実時間で画像を動かすには高価なビデオアクセラレータを必要とするなどの問題があった．

近年，JPEG画像圧縮/伸張を実時間で実行できる高性能かつ安価なハードウェアが提供されるようになってきた．本研究では，ハードウェアJPEGボードを利用して自然感の高い擬人化エージェント動画像の実時間生成を可能にした．さまざまな顔の向き，瞬き，口の形を変えた複数の顔画像をあらかじめJPEG形式に圧縮しておき，応答文やユーザの位置に合わせて必要な圧縮画像を実行時に任意の順序でつなぎ合わせ，ハードウェアJPEGボードを用いて実時間伸張する．

2. 擬人化エージェントシステムの全体構成

擬人化エージェントシステムの全体構成を図1に示す．システムは，対話処理部，ユーザ認識部，エージェント生成部から構成される．

・対話処理部　ユーザからの音声による質問，依頼に対する応答文の生成処理を行い，音声で答える．またエージェント生成部に口の形を決定するための情報を伝える．

・ユーザ認識部　エージェントがユーザの方向を向いて対話するように，ユーザの位置をエージェント生成部に伝える．

・エージェント生成部　対話処理部，ユーザ認識部からの情報をもとにして該当するエージェントの圧縮イメージをつなぎ合わせ，ハードウェアJPEGボードで実時間伸張処理を行いディスプレーに表示する．

ハードウェアは，シリコングラフィックス社のUNIXワークステーションIndyとJPEG画像圧縮ボード（Cosmo Compress）を用いた．音声合成装置は，NTTデータ社のしゃべりん坊を用いた．

3. JPEGによる顔画像の実時間生成

3.1 顔画像の生成

擬人化エージェントの顔は，向き，瞬き，口の形の3要素で構成する．

・顔の向き　顔の向きの範囲は通常の対面対称形式で不自然さが生じないように，上下方向に30°，左右方向に45°動かすことができるように設定した．また人間の顔は微妙に動いている点に着目し，エージェントが静止している場合でも顔に自然な揺らぎを加えている．会話の内容に応じて「うなずき（首を上下に振る）」，「否定（首を左右に振る）」などの簡単な動作ができる．画像処理や位置センサと組み合わせることにより，話しかけている相手の方向を向いて話すことも

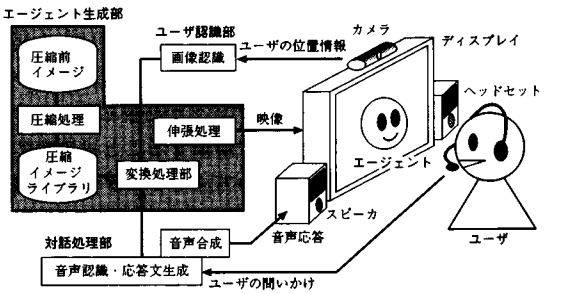

図1　擬人化エージェントシステムの全体構成
Fig. 1　System configuration.

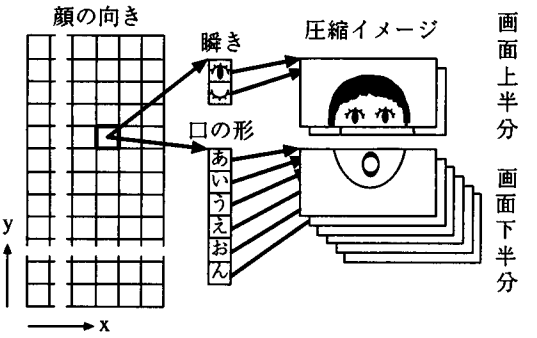

図2 データ構造
Fig. 2 Data structure.

可能である．

・瞬き 目を開けた状態と閉じた状態の2通りを表現する．瞬きのタイミングは，システムがランダムに決定している．

・口 閉じた状態の他に，発話に連動して5種類の母音の形を表現する．口の形は，対話処理部で生成された応答文から母音列を抽出している．

エージェントの原画像は，ビデオカメラから取り込んだ実際の人物の顔画像を使用する．これを3次元頭部ワイヤフレームモデルにテクスチャマップして，さまざまな向きや瞬き，口の形などの変形操作を加えた顔画像を合成し，JPEG形式に圧縮する．実行時に，ユーザ認識部からの情報をもとに必要な画像を任意の順序でつなぎ合わせ，ハードウェアJPEGボードを用いて伸長する．

3.2 データ量の削減

図2に，使用したデータ構造を示す．例えば顔の向きを上下方向に30°，左右方向に45°，上下左右3°間隔で動かす場合，必要なフレーム数は，21（上下）×31（左右）×6（母音+閉じた状態）×2（瞬きの有無）=7,812フレームとなる．幅640ピクセル×高さ480ピクセル，1ピクセル4バイト（RGBA，Aはピクセルの不透明度を表す）の画像データ量は，1フレーム当り約1.2Mバイトである．従って合計では約9.2Gバイト（1.2M×7,812）の膨大なデータが必要になり，すべてのデータをそのままの形で蓄積することは現実的ではない．そこで，次の3種の方法でデータ量の削減を行っている．

（1） 画像のJPEG圧縮

1フレーム約1.2Mバイトのデータに対して，JPEG圧縮を行った場合の圧縮率（Nominal compression ratio）と，得られるデータサイズの測定結果を表1に示す．（JPEGにパラメタとして与える圧縮率は名目的なもので，画像の内容により変動する．表1にみられるように，ここでの画像に対しては名目的圧縮率以上の圧縮率が得られる．）画質が劣化しないようにデータを1/10～1/20程度に圧縮する場合が多いが，VSAでは常に顔に揺らぎを与えており，実時間で再生することから，圧縮率を18とした．

表1 圧縮率と得られるデータサイズの測定値
（1.2MBのデータ10フレームの平均値）
Table 1 Compression ratio and measured data size.

Nominal compression ratio	Compressed data size / frame	Measured compression ratio (out / in)
15	54.7KB	1:23
16	26.1KB	1:47
17	21.1KB	1:58
18	18.3KB	1:67
19	16.3KB	1:76

（2） 画面分割

顔のそれぞれの向きに対して，口の形（6種類）と瞬き（2種類）の組合せは，フルサイズ（640×480）で12種類（6×2）必要となる．これに対して画面を，目を含む上半分と口を含む下半分に分け，表示時につなぎ合わせるとハーフサイズ（640×240）が8種類（6+2）で済むことになる．

顔の画面を上下2分割する代わりに，目を開いて口を閉じた状態を基本として，閉じた目とさまざまな口の形の部分画像を上書きする方法も考えられる．しかし，例えば上書きする部分の大きさを100×50ピクセルとすると約19.5Kバイトのデータが必要となり，かつ位置合せなどの処理も複雑となる．これはフルサイズの画像を圧縮率18で圧縮したデータ量よりも大きいことから，本システムではすべての画像を圧縮する方法をとった．

（3） ブロック化

プロトタイプシステムでは顔の向きを3°単位にすることにより，1°単位の場合に比べてデータ量を1/9に削減した．しかし，これでは顔の揺らぎも常に3°単位になる．そこで顔の揺らぎに変化をもたせるために，「あ」～「お」の発音に対応する口の形をした画面下半分の顔は，図3に示すように，3°単位のブロックの中心のまわりに1°ずつずれた方向を向くように配置した．すなわち，画面上半分の顔は基本的にユーザの位置に依存した各ブロックの中心を向いているが，下

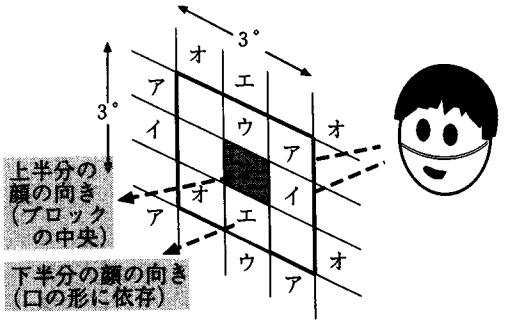

図3 1ブロック内での口の形状と顔画像の向き（上半分の顔は，基本的にユーザの位置により決まる3°単位のブロックの中央を向いている．下半分の顔の向きは，口形状で決まる．）
Fig. 3 Orientation of mouth images within one block.

図4 実 行 例
Fig. 4 A VSA image.

半分の顔の向きは上半分の顔の向きと口形状で決まる．口を閉じた場合（発音「ん」に対応），上半分の顔も下半分の顔も，両方ともブロックの中心を向く．発話している瞬間には画面上半分の顔の向きと下半分の顔の向きは1°ずれることになるが，口の形が連続的に変化するため，違和感はない．また発話終了時には口を閉じるため，画面上半分の顔の向きと下半分の顔の向きはもとのように同じになる．これにより，発話に連動して口を開くと顔自体が細かく揺らぐことになり，データ量は少なくとも自然感を高めることになる．

この結果，ブロックを上下左右に3°間隔にした場合，1ブロック当りの圧縮後のデータ量は約70Kバイトとなる．従って必要なデータ量は，システム全体で約45Mバイトとなる．このデータ量であるとシステム起動時にすべてのデータをメモリ上に展開可能で，JPEG伸長ハードウェアに切れ目なくデータを供給し，スムースな動画生成が可能になる．45Mバイト以上のメモリ容量はパーソナルコンピュータでも搭載されるものが増えてきているので，これによって自然感の高い擬人化エージェントインタフェースを実時間で動かすことができる．更に集積度の高いメモリが安価で大量に利用できるようになれば，本手法の拡張により表情をもつエージェント像についても同様に取り扱うことが可能である．実行例を図4に示す．

4. むすび

本研究では，利用が普及しつつあるハードウェアJPEG圧縮/伸張ボードを用いて，高価で特別なビデオアクセラレータを用いることなく，自然感の高い擬人化エージェント像の実時間生成が行えることを示した．銀行のATM装置や会社の受付システムなどの各種インフォメーション端末への応用にも有用と考えられる．

謝辞 本研究はNEDO提案公募型・重点分野研究開発事業，文部省科研費試験研究（No.06558045），および同奨励研究（No.08780246）の支援を受けた．

文 献

[1] H. Dohi and M. Ishizuka, "Realtime Synthesis of a Realistic Anthropomorphous Agent toward Advanced Human-Computer Interaction," Human-Computer Interaction: Software and Hardware Interfaces (Eds. by G. Salvendy and M. Smith), Elsevier, pp.152–157, 1993.

[2] 土肥 浩，石塚 満："WWW/Mosaicと結合した自然感の高い擬人化エージェントインタフェース，"信学論（D-II），vol.J79-D-II, no.4, pp.585–591, 1996.

[3] H. Dohi and M. Ishizuka, "A Visual Software Agent: An Internet-Based Interface Agent with Rocking Realistic Face and Speech Dialog Function," Working Notes of AAAI-96 Workshop on Internet-Based Information Systems, pp.35–40, 1996.

（平成7年12月25日受付，8年6月14日再受付）