ヒューマンインタフェース 80－5
音声言語情報処理 23－5
（1998. 10. 16）

ページエージェント：Webページからダウンロードする擬人化エージェント

土肥 浩　　石塚 満
東京大学工学系研究科電子情報工学専攻
〒113–8656 東京都文京区本郷7–3–1
TEL: (03)3812–2111 ex.6755
FAX: (03)5802-2990
E-mail: dohi@miv.t.u-tokyo.ac.jp

本論文では，Webページの作者がマルチメディア・コンテンツの一部としてインタフェースエージェントを記述し，Webページと連動して実用的な時間内でネットワークからダウンロードした擬人化エージェントを次々と切り替えることができるインタフェース，ビジュアル・ページエージェント VPA (Visual Page Agent) について述べる．VPAのエージェント・キャラクタは1枚の顔写真をもとにして生成される．VSAエディタを用いることにより，誰でも容易に動きのある顔画像キャラクタを作ることができる．このエージェントをWebページと関連づけることにより，ユーザがWebページを移動するたびに，そのページに関係するエージェントが画面上に現われてガイドしてくれる．エージェントはメッセージを伝えたり，またユーザと簡単な音声対話ができる．クライアント側のユーザにとっては，情報発信者の顔が見えるというメリットがある．

Visual Page Agent: Network–downloadable Anthropomorphic Interface Agent

Hiroshi Dohi　　Mitsuru Ishizuka
Dept. Information and Communication Engineering,
School of Engineering, University of Tokyo

This paper describes a network–downloadable anthropomorphic interface agent with a realistic face, called VPA(Visual Page Agent). An author can assign the facial image and some properties of the interface agent to own web page as the part of the multimedia contents. Whenever a user opens the web page, the agent with assigned face is downloaded and then appears on a display. The agent equips a simple speech dialog function, therefore it delivers author's messages to the user and can reply simple question.

1 はじめに

ユーザとのインタラクションや情報のプレゼンテーションのためにキャラクタを用いる擬人化エージェントインタフェースが注目されている．親しみがあり動きのあるキャラクタは，機械とのインタフェースにおける心理的抵抗感を和らげる．生き生きとした動きをする人間や仮想生物のキャラクタは，"Life-like Agent"と呼ばれる．これらのエージェント・キャラクタはユーザからの入力に反応したり，さまざまな表情や姿勢，動作をすることができるように作られている．ただし魅力的な，動きのあるエージェント・キャラクタを作ることは容易ではなく，使用できるキャラクタの種類は通常，1体～数体に限られていた．

本論文では，Webページの作者がマルチメディア・コンテンツの一部としてインタフェースエージェントを記述し，Webページと連動して実用的な時間内でネットワークからダウンロードした擬人化エージェントを次々と切り替えることができるインタフェース，ビジュアル・ページエージェント VPA (Visual Page Agent) のプロトタイプについて述べる．VPA のエージェント・キャラクタは1枚の顔写真をもとにして生成される．VSA エディタを用いることにより，誰でも容易に動きのある顔画像キャラクタを作ることができる．このエージェントをWebページと関連づけることにより，WWW ページを移動するたびに，そのページに関係するエージェントが画面上に現われてガイドしてくれる．そのページを紹介したりメッセージを伝えたり，またユーザと簡単な音声対話ができるエージェントの顔やエージェント属性をWebページの作者がWWW サーバ側で指定できる．クライアント側のユーザにとっては，情報発信者の顔が見えるというメリットがある．

2 ビジュアル・ページエージェント

2.1 VSA & VPA

VPA は，我々がこれまで開発してきた擬人化エージェント VSA(Visual Software Agent) をベースとしている．VSA は，自然感の高い人間の顔を持ち，インタラクティブにユーザと対話できる知的擬人化インタフェースエージェントである．人間の日常的な対面型 (face-to-face) コミュニケーションスタイルに基づいて，顔写真をもとに CG 合成されたエージェントとテレビ電話でおしゃべりするような感覚で，操作方法を覚えたり練習したりすることなしに，誰もが高度な計

図 1: VSA/VPA interface connected with Netscape

算機の能力をフルに活用できるインタフェースの実現を目指している．VSA はすでに Netscape や Mosaic と結合されており，これらの WWW ブラウザ画面を通してさまざまな情報をユーザに提供する [3][4]．これにより VSA で提供するコンテンツデータを，HTML 言語をはじめとして WWW で標準的に使用されているフォーマットで記述できるだけでなく，これまでに WWW 上で蓄積されてきた膨大なマルチメディアデータを WWW と同じように VSA から利用することができる．ユーザは，通常のマウスやキーボードによる操作に加えて，アンカー文字列を発話するなど，VSA との簡単な音声対話によりインターネット情報空間を散策できる．また，電子メイルの到着や WWW から自動的に取得した最新の天気予報などを，音声で聞くこともできる．マウス／キーボードによる操作と音声による操作は同じ優先度をもつため，ユーザは状況や環境にあわせて，いつでも最適なインタフェースを選択できる．

VPA と VSA はプログラムの多くの部分を共有しており，Webページと連動して顔が次々と切り替わる点を除いてシステムの外観は同じである．以下，擬人化エージェントインタフェースの基本部分を VSA，顔が切り替わる部分（及びその機能をもつシステム）を VPA と呼んでいる．

2.2 顔が替わるインタフェース

擬人化エージェントシステムには，親しみがあり動きのあるキャラクタが使われている．これらのシステムでは，2D や 3D のアニメーション CG を利用しているものが大勢を占めている．ただし魅力的なエージェント・キャラクタをデザインするにはセンスが必要であり，またそれに表情や動きを与える作業は容易ではない．一般に，使用できるキャラクタの種類は 1 体～数体に限られている．

WebPersona[1] では，エージェントは 1 体である．プロのアーティストが描いた 200 フレーム以上の多彩なアニメーションを使用している．

GazeToTalk[2] では，男女 2 体のエージェントがあり，身振りを伴う多彩な表情をアニメーションで提示する． 動作パターンは，“NORMAL”，“HAPPY”，“SORRY” など，男女各 15 種類が用意されている

Microsoft Agent[5] では，「Merlin（魔法使い）」，「Genie （妖霊）」，「Robby（ロボット）」の 3 体が標準で用意されている．それぞれ「うなずく」「手を振る」「中立姿勢に戻る」など約 100 種類の動画（[5] によれば平均的な動画は 14 フレーム程度）で構成されており，データ量は 1 キャラクタ当り約 3MB である．また，ユーザが独自のキャラクタを追加する手段として Agent Character Editor を提供している．

いずれのシステムもユーザが実際に新しいオリジナルキャラクタを追加することは困難であり，たとえ追加出来る場合でも任意のキャラクタに次々と切り替えていくためには，あらかじめローカル・システムにデータを用意しておかなければならない．つまり，従来の擬人化エージェントインタフェースでは，あらかじめシステムに用意されているキャラクタを利用することが事実上の前提となっている．これらはツアーガイドのように同じエージェントがいつもユーザと一緒に行動しながら，質問に答えたり助言をしてくれるインタフェースであるといえる．

これに対して，同じようにユーザの質問に答えたり助言するインタフェースには，もう一つの形態があることがわかる．エージェントはユーザと行動をともにするのではなく，それぞれの持ち場でユーザの到着を待ち，ユーザが来れば応対する方法である．各エージェントにはそれぞれ自分の専門領域があり，その分野の内容についてエキスパートである．ただし専門外のことについては，エキスパートであるかどうか分からない．エージェントの姿は，それぞれ異なる．

VPA システムでは，顔写真を使って容易に新しいエージェント顔画像を追加する手段を提供する．正面から撮影した 1 枚の顔写真を 3 次元頭部ワイヤフレームモデルにテクスチャマッピングすることにより，動きのあるエージェント顔画像を生成する．このキャラクタ顔画像は瞬きをしたり，廻りを見渡したり，合成音声に合わせて喋るように口を開いたりする．

2.3 Web との連動

この顔写真，エージェント属性，および音声対話用データを WWW サーバ側に置き，必要に応じてダウンロードする．Web ページと関連づけることにより，ページ毎にエージェントの顔を自動的に切り替えることができる．

情報発信者側からみれば，エージェントをマルチメディア・コンテンツの一部として記述できることになる．ユーザは，情報発信者の「顔」を見ることができる．現在，数多くのページに顔写真が載せられている．単に顔写真を含んだ文章が画面に現れるのに較べて，動きのあるエージェントが画面に現れてメッセージを語りかける方がユーザに強く印象づけられるであろうことは容易に想像できる．さらにエージェントからユーザへの一方向のメッセージ伝達ではなく，ユーザがそのエージェントに対して質問し，エージェントがそれに答える双方向のコミュニケーションを実現することができれば，より魅力的なインタフェースとなる．

例えば企業の「社長あいさつ」の Web ページを開けば，それに連動して Web ブラウザの隣りにいる擬人化エージェントが社長の顔に自動的に替わり，そのページの内容をしゃべったり簡単な音声による質問に答えたりすることができるようになる．

2.4 3 点マッチング法

顔写真を 3 次元頭部モデルにテクスチャマップするためには，両者の対応点のマッチングをとることが必要になる．普通はできるだけ多くの特徴点を正確に合わせようとするため，非常に手間と時間のかかる作業となる．そこで，特徴点を重要度の高い右目，左目，口の 3 点に絞りこむことにより，マッチングを素早く行うことができる 3 点マッチング法を提案する．

瞬きをしたり口を開くといった変形操作をするためには，これらの 3 点が正確に対応していることが重要である．もしこれらの対応がずれていると，音声に合わせて顔の一部が裂けるといった事態になりかねない．これに対して，それ以外の点のマッチングはそれほど

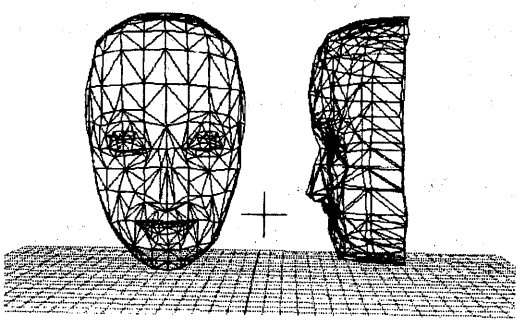

図 2: 3D face model

重要ではない．何故ならテクスチャマップの原理は，原画像の物体のテキスチャ（色情報）を変換後の物体の対応する位置に，ピクセル単位で転写することである．そのため，それほど正確に対応していなくても，不自然さは目立たない．

入力画像として，正面から撮影した1枚の顔写真を使用する．わずかに顎を引いて正面を向き，口を閉じた，表情のない顔を仮定している．正面を向くとは，カメラの軸廻りには回転してもよい（例えば，首をかしげてもよい）が，上下や横を向いたりはしないことをいう．

両目の位置，及び口の中心の位置が分かると，システムはその位置関係から3次元頭部モデルが入力画像にうまく重なるように，移動量，スケール，カメラの軸廻りの回転角度などの変換パラメタを計算する．これには，わずかに顎を引いて正面を向いた時，両目を結ぶ線は顔画像の高さ（見かけの頭頂部から顎まで）のほぼ1/2のところを通ることを利用した．これにより，画像から切り出す顔のテクスチャの位置が決まる．またこの情報を利用して，エージェントを合成する際に顔が画面中央に，適当な大きさで表示されるようにする．頭部モデルの輪郭を構成するエッジは，実際の顔画像よりも一廻り小さめになるように設定されている．口を開いたときに見える歯のテクスチャは，別途用意してある．

2.5 VSA Editor

VSA Editorは，3点マッチング法を利用してビデオキャプチャした顔画像と3次元頭部モデルとのマッチングをとるソフトウェアである．また，VPAインタフェースで必要となるエージェント属性ファイル（拡張子 ".vsa"）とJPEG圧縮した顔画像ファイルを生成する．

顔画像と頭部モデルのマッチングには，入力画像を取り込み，その顔画像の3特徴点の上を順番にマウスでクリックするだけでよい．必要があれば，一つ一つの頂点をマウスでドラッグして動かすこともできるので，十分な時間をかけることができる場合には詳細な合わせ込みをすることも可能である．このとき，顔の奥行き方向の値は変更されない．通常，画面上の3特徴点の上でマウスをクリックし，その廻りのいくつかの頂点を動かすだけであれば，カメラで顔画像を撮影してからエージェントが合成されるまでに1分もかからない．

画像認識の分野では，顔画像から目や口などのパーツを抽出する手法が数多く研究されており，良好な結果が得られている．したがって本手法と組み合わせれば，右目，左目，口の三つの特徴点を抽出し，頭部モデルとのマッチングを人手を介さずに自動化することも可能である．また肌の色情報等を使って，顔の輪郭エッジをより正確に，自動的に合わせ込むことも可能であろう．

3 実装

3.1 VPA タグ

WWWサーバ側で，自分の好きな顔の擬人化エージェントを任意のWebページと関連づけるために「VPAタグ」を導入した．VPAタグは，次に述べるエージェント属性ファイルのURLを含んでいる．これをHTMLファイル中に，1行記述するだけでよい．

<VPA VSA="属性ファイルのURL" />

もしクライアントがVPAシステムであれば，Netscape上のマウス操作あるいはエージェントとの簡単な音声対話により，VPAタグを含んだWebページを開くとNetscape画面にその内容が表示され，それに連動して自動的にエージェントの顔が切り替わる．エージェントの顔を替えるには単にそのページを開くだけでよく，ページが表示された後に特定のアンカーをクリックする必要はない．

通常のWebブラウザで直接，VPAタグを含んだWebページを開いても全く問題は生じない．VPAタグは，未知のタグとして単に無視されるだけである．この場合，VPAが設定されていない普通のWebページと同じに扱われ，Webブラウザにはその内容が表示される．

3.2 エージェント属性ファイル

エージェント属性ファイルは拡張子 ".vsa" を持つテキストファイルで，以下の項目を含んでいる．

- 顔画像ファイル（JPEG 圧縮）の URL
- 言語（日本語／英語）
- 音声属性（男声／女声，ピッチ，抑揚等）
- メッセージテキスト 1（初回）
- メッセージテキスト 2（2 回目以降）
- そのページに特有の（アンカーリスト以外の）音声キーワード－ URL 対応テーブル
- 顔形状データ

メッセージテキストは，初めてそのエージェントが現れる（選択された）ときに伝えるメッセージ 1 と，Netscape の Back ボタン等でそのページに戻ってきたときに話すメッセージ 2 の 2 種類が用意されている．例えば，初回には比較的長いメッセージを伝え，2 回目以降では名前だけを話すように設定できる．

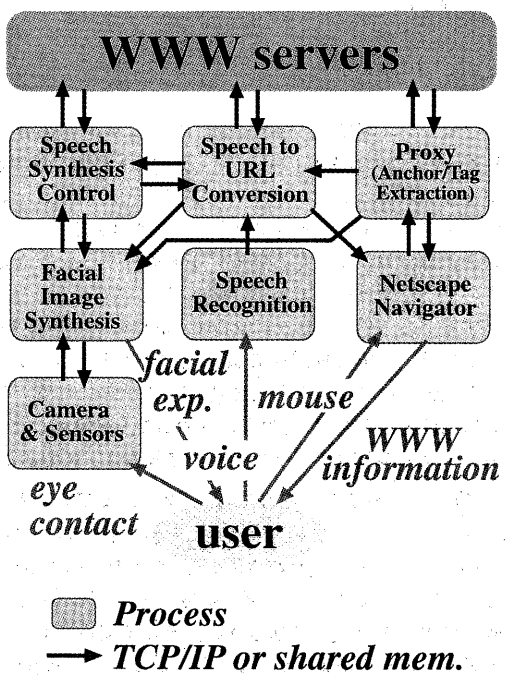

図 3: VPA implementation

3.3 システム構成

VPA は VSA インタフェースシステムをベースにして，従来の WWW の枠組みの上に実装されている．顔画像の実時間生成，連続音声認識，規則音声合成，基本的な音声対話用データなどの基本機能はすべてクライアント側に存在する．顔画像，エージェント属性，および付加的な音声対話用データ等を WWW サーバから取得する．http サーバや Web ブラウザには一切手を加えていないので，稼働中の WWW システムがそのまま使える．

VPA インタフェースシステムは，Netscape を含めて六つ，あるいはそれ以上のプロセスで構成される．プロセス間は，TCP/IP（一部，共有メモリ）で接続されている．基本的な構成は VSA システムと同じであるが，VPA では同時に複数のコネクションを張ることができ，より複雑なプロセス間通信を行えるようになっている．各プロセスは，複数のワークステーション上で，並行あるいは並列に動作する．これにより，例えば Netscape がデータを転送中でも，エージェントが顔を動かしている間でも，ユーザは任意の時点で音声入力することが可能である．

1. VSA Proxy プロセスは VPA タグを検出すると，エージェント属性ファイルの URL を顔画像生成プロセスに転送する．
2. 顔画像生成プロセスは，そのデータがすでにダウンロードされ，クライアント側にキャッシュされているかどうかをチェックする．
3. もし初めてのエージェントであれば，Netscape のページアクセスと並列に WWW サーバからエージェント属性ファイルをダウンロードする．さらに，エージェント属性ファイルに記述された顔画像ファイルを WWW サーバからダウンロードする．
4. 伸張した顔画像ファイルと顔形状データから，擬人化エージェント顔画像を生成する．
5. 規則音声合成装置に，エージェントの音声属性を設定する．
6. 音声キーワード－ URL 対応表を，インタフェースマネージャ（音声キーワード－ URL 変換プロセス）に転送する．
7. 規則音声合成装置と連動して，エージェントがテキスト 1（初回）あるいはテキスト 2（2 回目以降）をユーザに話す．

4 評価

4.1 ネットワーク・オーバヘッド

プロトタイプでは，640×480×24ビット（RGB各8ビット）の顔画像を使用した．JPEG圧縮後の顔画像ファイルのサイズはバイナリで約30KB，エージェント属性ファイルのサイズはテキストで約15KBである．この大部分は，3次元頭部モデルを構成する約500頂点のxyz座標データである．従って，1人のVPAを生成するためにダウンロードするネットワーク・オーバーヘッドは，JPEG圧縮後の顔画像ファイルとエージェント属性ファイル，合わせて約40〜50KBになる．これは画像1〜2枚分程度の大きさであり，実用上，十分許容できる時間内にダウンロードできるサイズである．

HTMLファイルからのVPAタグ，およびアンカーリスト（アンカー文字列とURLの対応表）の抽出はProxyプロセスで行われる．VPAタグを検出すると，そこに含まれるエージェント属性ファイルのURLが顔画像生成プロセスに転送され，そこで属性ファイルのダウンロードが実行される．したがってNetscapeで表示されるデータのダウンロードと，VPA用ファイルのダウンロードは並行，あるいは並列に実行される．

4.2 キャッシュ

顔画像ファイルとエージェント属性ファイルの内容は，それぞれ独立に顔画像生成プロセスでキャッシュされる．メッセージテキスト1/2の選択（初回であるか2回目以降であるか）は，データがキャッシュされているかどうかで判断している．手動で，そのキャッシュエントリを無効にする（強制的にメッセージテキスト1を選択する）こともできる．

このうち顔画像ファイルのキャッシュは2レベルあり，LRUアルゴリズムにより最近使われた数枚は伸長した状態で，それ以外はJPEG圧縮した状態でキャッシュされる．例えば"Back"ボタンや『前のページを見せてください』という発話により一つ前のページに戻る場合，属性ファイルも顔画像ファイルもキャッシュにヒットするのでネットワークオーバーヘッドはほぼ0である．また顔画像ファイルのJPEG伸張も必要ないので，一瞬でエージェントの顔が切り替わる．音声属性も，顔に合わせて切り替わる．ただし現在は音声合成装置の制約により，日本語の音声属性の種類は限られている．

5 まとめ

エージェントとの音声対話によりインターネット情報空間を散策できるVSAインタフェースシステムをベースとして，Webページ毎にエージェントの顔が自動的に切り替わるビジュアル・ページエージェントVPAのプロトタイプを実現した．

Webページの作者が，マルチメディア・コンテンツの一部として，そのページを紹介したりメッセージを伝えたり，またユーザと簡単な音声対話ができるインタフェースエージェントを記述できる．情報発信者はインパクトのある魅力的なインタフェースを提供でき，ユーザは情報発信者の顔を見ることができる．新しいネットワーク型情報媒介機構として，さまざまな応用が期待される．

参考文献

[1] E. André, T. Rist and J. Müller: "WebPersona: A Life-like Presentation Agent for the World-Wide Web", *Proc. IJCAI-97 Workshop on Animated Interface Agents: Making Them Intelligent*, pp.53–60, 1997

[2] 知野，福井，山口，鈴木，田中: "GazeToTalk: メタコミニケーション能力を持つ非言語メッセージ利用インタフェース", *Interaction98*, pp.169–176, 1998

[3] H. Dohi and M. Ishizuka: "A Visual Software Agent: An Internet-Based Interface Agent with Rocking Realistic Face and Speech Dialog Function", *AAAI tech. report 'Internet-based Information Systems', WS-96-06*, pp.35–40, 1996

[4] H. Dohi and M. Ishizuka: "Visual Software Agent: A Realistic Face-to-Face Style Interface connected with WWW/Netscape", *Proc. IJCAI-97 Workshop on Intelligent Multimodal Systems*, pp.17–22, 1997

[5] Microsoft Corp.: "Developing for Microsoft Agent", Microsoft Press, 1998

[6] T. Oren, G. Salomon, K. Kreitman, and A. Don: "Guides: Characterizing the Interface", *The Art of Human-Computer Interface Design (B. Laurel eds.)*, Addison-Wesley, pp.367–381, 1990