手形状の認識に基づき実時間で反応する人物表情の動画像合成[†]

長谷川 修*・李 七 雨*
ウィワットウォンフラウィパット*・石 塚 満*

Real-time Moving Image Synthesis of Human Facial Expressions Based on Hand Shape Recognition

Osamu HASEGAWA*, Chil-Woo LEE*,
Wiwat WONGWARAWIPAT* and Mitsuru ISHIZUKA*

To realize a human-like software-robot for better human interface system, we must integrate the technologies of computer graphics, computer vision, speech recognition, voice synthesis and others. We have named this type of the software robot to be served in an advanced human interface system as "Visual Software Agent (VSA)", and begun our research project for its realization. In the VSA, it is very important to realize real-time processing of several tasks to achieve the real-time response of the software robot. In this paper, we describe a real-time image recognition and image synthesis system, "TN-VIT (Transputer Network with Visual Interface for Transputers)", as a first step to build the "VSA". In this prototype system, a human face (agent) is shown on the CRT, and he/she gazes at operators and changes his/her facial expressions in response to operator's hand shape in real-time. Therefore, the system recognizes hand shape of an operator, then synthesizes human facial expressions in real-time. The VIT is our original Transputer board with functional 32-bit parallel visual data bus. The maximum visual data transfer speed of our VIT is 100 Mbyte per second. The TN-VIT system consists of 16 VITs, 16 normal Transputers, 1 personal computer, 1 CCD camera, 1 A/D board and 1 D/A board. The 16 VITs are connected as a ring-type bus, and 2 of them are used for image recognition and the rest of them are used for image synthesis. Experiments show that the TN-VIT system can synthesize natural moving human face images and his/her emotional expressions at 8 frames per second in response to the hand shape of an operator. Details of the hardware and software are described in this paper along with experimental results including some samples of original image and synthesized images.

Key Words: human interface, software robot, parallel processing, real-time image recognition and synthesis, hand shape reading

1. はじめに

ヒューマンインタフェースや人工現実感、ソフトウェアロボットなどの重要性は近年広く認識され、現在内外の各機関で活発に研究・開発が進められている^{1),2),8)}。一方でコンピュータの処理能力は年々大幅な進歩を続けており、このハードウェア技術を主とした進歩を人間とコンピュータとの良い接面^(*)1)の実現のために活用するこ

とは非常に望ましい。特にヒューマンインタフェースなどにおいて重要な、画像合成や画像処理、画像理解の高速化のために処理の並列化は今後の重要な方向とされ、並列コンピュータの利用とその性能の追求、また並列アルゴリズムを含む技術の確立が求められている^{3),4)}。われわれはこのような背景に基づき、近未来の情報機器のヒューマンインタフェースとして重要な役割を果たすと考える、自然な人間の姿を有して実時間で動作し、限定はされているがコミュニケーション機能を有する「ビジュアル・ソフトウェアエージェント (Visual Software Agent, 以下 VSA)」の概念を提案し、並列コンピュータ

[†] 第7回ヒューマンインタフェースシンポジウムで発表 (1991・10)

* 東京大学生産技術研究所 東京都港区六本木 7-22-1

* Inst. of Industrial Science, University of Tokyo, Minato-ku, Tokyo

(Received September 19, 1991)

(Revised June 17, 1992)

(注1) 「接面」とは面的な広がりをもつ(接点でない)インタフェースのこと。インタフェースに関する認知科学の分野では使われることがある。

ネットワーク上に実現する研究を進めている⁵⁾。

本稿では、VSA の実現のために、自然な人の姿を有したモデルが実時間で入力画像を認識し反応する現状ではほかに例をみないシステムを、画像処理・合成用コンピュータ“TN-VIT (Transputer Network with Visual Interface for Transputers, 以下 TN-VIT)”上に作成したので、関連するハードウェア構成とソフトウェア・アルゴリズムを述べる。本システムによれば、カメラからユーザの手の挙動を画像入力し、その入力内容に応じ人物モデル(エージェント)が首を振る動作や表情が変化する過程を、実時間(毎秒約 8 コマ)で合成することができ、この人物モデルは、3 次元ワイヤフレームモデルに実際の人物像をテクスチャマッピングして作成した、このため本システムで用いた人物モデルは、ワイヤフレームを単にレンダリングしたモデルと比較して、非常に自然感のあるモデルとなった。また TN-VIT では、並列コンピュータの要素プロセッサとして 4 本の通信リンクを具備したトランスピュータ(英国 Inmos 社製)を用いたが、トランスピュータの通信リンク速度は現行で最大 20 Mbit/s と画像データの転送には不十分であることから、トランスピュータのローカルメモリに直接データ入出力でき、最大 100 Mbyte/sec のデータ転送能力をもつ独自の 32 bit 並列高速ビジュアルデータバス付きトランスピュータボード“VIT (Visual Interface for Transputers, 以下 VIT)”を 16 台作成し利用した。

2. ビジュアル・ソフトウェアエージェント(VSA)

本章では、われわれが画像認識・画像合成・知識ベース・並列処理の各技術を統合した新たなヒューマンインタフェースの概念として提案し研究を進めている、ビジュアル・ソフトウェアエージェント(VSA)の概要と構想について述べる。

Fig. 1⁶⁾ は VSA の実現のためのシステム構成案を示したものである。Fig. 1 に示すように、VSA は人間にとり親しみやすい種々の入力媒体(音声、静止・動画像など)をもち、それらを有機的に結合するための知識処理機能を備えた、統合型ヒューマンインタフェースである。

情報の視覚化は、人間に対する情報の伝達にはきわめて有効な手段である。ここでいう情報の視覚化には、単に静止画像や文書だけではなく、3 次元動画像も含んでいる。VSA では、一種のソフトウェアロボットともいえる、リアリティをもった実時間 3 次元動画像(人物像)をエージェントとしコンピュータとユーザの接面とする。このようなアプローチは、必要とされるコンピュータ能力などの点から、従来のヒューマンインタフェースにはあまりみられず、VSA の大きな特色である。このほかに

Fig. 1 A Plan of VSA System

も VSA には、以下に示すような特色をもたせる。

(1) 高速な応答 人間とコンピュータの良いインタフェースには、ビジュアル情報も含めた十分な応答速度、遅滞感のない迅速なコミュニケーションが不可欠である。VSA では並列トランスピュータおよび VIT(後述)を用い、高速な処理と応答を実現する。

(2) 自然に近いコミュニケーション、操作感

VSA では実際の人物画像を基に、自然に近い人物動画像を合成してユーザとコンピュータとの接面とし、画像入力や音声入出力も加え自然感のあるコミュニケーションの実現を目指す。

(3) 豊富な情報の提供方法・手段 通常ユーザに効率的・効果的に情報を提供するために、システムには豊富な情報の提供方法・手段が求められる。VSA ではユーザへの情報の提供において、3 次元人物動画像のほかに、文書・静止画像・動画像などを目的や状況に応じて選択可能とする。

3. VIT と TN-VIT

本章ではユーザの挙動の認識と人物像の実時間合成のために用いた、VIT⁷⁾の概要および TN-VIT について述べる。

3.1 VIT

画像認識と画像合成の高速化のために、VSA では並列トランスピュータ・ネットワークを採用したが、現状ではトランスピュータのリンクの通信速度は画像データの転送には不十分である¹⁰⁾。そこでわれわれはトランスピュータのローカルメモリに直接データ入出力でき、ビデオレートで画像を転送できる 32 bit 並列高速ビジュアルデータバス付きトランスピュータボードを開発し、VIT と名付けた。各 VIT は、T 800 トランスピュータ(1.5 MFLOPS, 10 MIPS)⁹⁾を 1 台、2 MB のプログラムメモリ、1 MB のローカル画像メモリを有する。通信系では、通常の 4 本のシリアルリンクのほかに、32-bit 画像

Fig. 2 Function of visual data bus of VIT

データバスを有し、画像のデータ転送速度は、最大 100 Mbytes/sec (40 nsec/画素)である。さらに VIT では、通常のトランスピュータの開発環境 (TDS, OCCAM など)をそのまま利用することができ、プログラミングは容易である。

また各 VIT は、カメラから画像データバスに入力された画像データの処理、およびローカルメモリへの描画が可能である。画像の表示時には、VIT は隣の VIT から画像データとみずからのローカル画像メモリ中のデータを、六つの画像合成モード⁷⁾から選択してピクセル単位で合成し、出力することができる。VIT の画像データバスは、マルチプロセッサにおける読み出しだけの共通メモリと同様に働き、どの VIT でも高速にバスにアクセスできるが、アクセスの競合問題は起こらない。これは Fig. 2 において、画像データは方向性を有する 32 ビットパラレル画像バスを通じ、図の左側の VIT から順に 1 度だけ通ることによる。すなわち全ての VIT は、画像バスによりアレイ状に結合されている。またこれとは別に、VIT はほかの VIT とシリアルリンクを通じて通信することができる。ここで各 VIT において画像データの入出力に要する時間は、1 画素クロック (ノン・インターレースモードで 33 ns, インターレースモードで 66 ns) であり、ほぼ無視できる。実際に作成された VIT は、ダブルハイト・ユーロカードの大きさであり、23.25×22.00 cm である。

3.2 TN-VIT

TN-VIT とは、VIT を中心とした並列トランスピュータネットワークと画像の入出力のための周辺装置の総称である。すなわち TN-VIT は、VIT の各種機能を活用し、画像処理と画像合成が同一のハードウェア上で行えるシステムであり、さらに通常のトランスピュータによる知識ベースなどを付加すれば、高度な画像の認識・推論処理が可能となる。現在、VIT が 16 台、通常のトランスピュータ 24 台が利用可能となっている。これらの構成は、ト

Photo. 1 Outline of TN-VIT

ランスピュータの通常のリンクを活用することにより任意に設定が可能であり、目的に応じた編成で使うことができる。また VIT では、画像処理に用いるか、または画像合成に用いるかの選択をソフトウェアで動的に行うことができるため、TN-VIT の 16 台の VIT においては、画像処理部と画像合成部の構成は処理対象に応じ、動的に変更することも可能である。このうち、本稿で述べるユーザの手の挙動による入力画像の解析および顔画像の合成には、16 台の VIT のみを用いた。16 台の負荷分散方法については後に詳しく述べる。Photo. 1 に TN-VIT システムの概観を示す。現行の VIT の数 (16 台) は、人物モデルの画像処理において、これを動画像として知覚するために必要な最少の台数を算出・予測して設定した。しかし現在、VIT の台数の拡張作業を進めており、近い将来に 32 台とする予定である。

本システムに同一または類似の機能を有するシステムとしては、現状では国の内外に報告がみられない。たとえば高速画像合成マシンとして代表的なシルコングラフィックス社の IRIS は、高度な画像合成機能を有するが画像処理機能の付加は困難であり、本システムとは基本的に設計方針が異なる。

4. 画像処理によるコマンドの入力

4.1 手の形状識別の利用

手の形状識別を利用したシステムは、古くは MIT メディアラボの “Put That There” システム (磁気センサ利用、音声入力併用)¹⁰⁾ が代表的であり、近年も研究が盛んである^{11), 12)} が、その多くがデータ入力にデータグローブなどの特殊なハードウェアを用いている。

これに対し本システムでは、ユーザの手の挙動をカラー画像入力し、これに人物モデル (エージェント) が反応する。具体的には、手の位置に加え指の数に応じた 1 から 5 までの五つのコマンドを入力とし、人物モデルはユー

ザの手の位置を、顔を振ることによりトラッキングすると同時に、提示された指の本数により随時表情を変化させることができる。またユーザが手を握り拳とすれば、手形状の認識処理をオフにすることができる。これにより、必要があれば、手を握り拳としてコマンド入力を行わず、画像入力用カメラの前で手を移動させることも可能である。ここで、ユーザの手の挙動の入力は、データグローブなどのような特殊なハードウェアは用いず、画像処理のみにより行うこととした。これはヒューマンインタフェース研究の目的を考える際、コマンド入力が簡易であることが必須であり、その点手の形状識別は誰にでも容易に利用可能である点に着目したことによる。これは本システムの大きな特色の一つである。

画像処理のみによる手形状の識別の類似の研究としては、石淵ら¹³⁾の研究および福本ら¹⁴⁾の研究がある。石淵らの研究では、ステレオ画像により手の空間位置の測定および提示された指の同定が行えると報告されている。しかし本システムと比較すると、システム構成は複雑であり、処理速度も約 2 秒(著者の口頭による説明による。本システムでは 1 台の VIT で約 0.5 秒、以下括弧内は本システムにおける状況)を要する。また処理開始時に初期設定が必要である点(不必要)や、半袖着用時に処理ができない(可能)などの問題点も報告されている。一方福本らの研究では同様に 2 台のカメラを用いるが、手の 2 次元平面内での形状識別のみを行い、3 次元的な空間位置の測定は行っていない。また補助光源として赤外線照明が必要であり(通常の室内蛍光灯環境下で実行可能)、システム構成が複雑である。また処理開始時に初期設定が必要(不必要)であり、ユーザが変わる度に設定し直す必要があるなどの問題点がある。

以下に手の形状識別利用の場合の長所と短所を、データグローブ利用の場合と比較して列挙する。

[長所]

- ① 入力のための特別な装置を必要としない。
- ② 右手用、左手用などの区別がない。
- ③ 手の大きさ(個人差)に影響を受けない。
- ④ 手の 2 次元平面内での移動距離が容易に計測可能。
- ⑤ 処理開始時の初期設定などの必要がない。

[短所]

- ① 環境(照明の位置など)に影響を受ける。
- ② 掌または手の甲を、常にカメラに向けている必要がある。
- ③ 各指の関節ごとの曲げ角は計測不可能。
- ④ 指の識別が困難。
- ⑤ 現段階では背景を調整する必要がある。本システ

ムでは背景を白色とした。

- ⑥ 手の 3 次元空間内での移動距離は計測困難。

[その他]

このほかにも、本システムで用いた手の形状識別の読みとりアルゴリズムには以下のような特長がある。

- ① 手の回転角に影響を受けない。つまり、掌か手の甲がカメラに向いていれば、どの向きでもデータ入力が可能。
- ② 手袋などを装着する必要はなく、また長袖でも半袖でも読みとり可能(ただし長袖の場合、袖のしわの状態、柄によっては認識率が低下する)。
- ③ 必要ならば、一度に両手を用いることも可能。

4.2 手の形状識別の認識アルゴリズム

本節では、手の形状識別の認識(提示された手の位置と指の数のカウント)アルゴリズムを示す。なお下記の画像処理過程は、すべて VIT 上で行っている。

- ① 手の画像をカラー RGB 入力する (512×400 pixel)。
- ② ローパスフィルタによる画像の平滑化を行う。
- ③ 画像の x, y 方向それぞれ 10 ピクセルごとにデータをサンプリングすることにより、画像の圧縮を行う。
- ④ カラー RGB 入力した Red : Green : Blue の比が手の画像のそれらの比に近く、かつある程度以上の明るさをもつデータを、手で示すデータとして抽出する。
- ⑤ ④で抽出されたデータに対し、5×5 サイズで計測して④で抽出されたデータを 10 以上、16 以下含む場合、その中心データを対象データとして抽出する。これを画像全体にわたり行う(この操作をここでは、5×5 サイズの検索と呼ぶことにする。以下の⑥・⑦においても同様)。これは主として、指の部分、および掌または手の甲の輪郭線を抽出することに等しい。Fig. 3 に指の部分の抽出処理過程の例を示す。
- ⑥ ⑤で抽出されたデータに対し、7×7 サイズで計測して④で抽出されたデータを 21 以上、35 以下含む場合、その中心データを対象データとして抽出する。

Fig. 3 5×5 size data extraction

Fig. 4 7×7 size data extraction

これを画像全体にわたり行う(この操作をここでは7×7サイズの検索と呼ぶ)。これは⑤で抽出されたデータ群のうち、掌または手の甲でないか、またはそれらから遠い部分、すなわち指先を抽出することに等しい。Fig. 4に指の先端部分の抽出処理過程の例を示す。

⑦ ⑥で抽出されたデータに対し、3×3サイズで計測して近傍に存在するデータのグルーピングを行う(この操作をここでは3×3サイズの検索と呼ぶ)。これは⑥までの過程で抽出された、同一の指の先端に対する複数のデータ群を一つにまとめることを意味する。

以上のアルゴリズムにより、指の先端のみを抽出することができる。システムはこの抽出されたデータの数数を数えることにより、カメラの前に提示された指の数数を認識することができる。また手の位置は、抽出されたデータ(群)の重心を求めることにより算出している。

5. 人物モデル(エージェント)の合成出力

本章では、本研究で用いた TN-VIT の構成、および人物モデルの実時間トラッキング・合成法について述べる。

5.1 システム構成

本研究で用いた TN-VIT の構成を Fig. 5 に示し、またその内訳を以下に示す。VIT: 16 台、パーソナル(ホスト)コンピュータ: 1 台、CCD カメラ: 1 台、A/D ボード: 1 台、D/A ボード: 1 台。

パーソナルコンピュータ(NEC PC-9801 RA)はトランスピュータネットワークのホストコンピュータとして用いる。ホストコンピュータは、簡易 OS である TDS 2 によるプログラムのエディット、プログラムのトランスピュータネットへのダウンロード、ファイルの管理を行う。合成の対象となる人物像の基本カラー画像(テキストチャデータ、以後人物画像とする)は、CCD カメラ・A/D ボードを通じ、TN-VIT に入力される。ここで人物画像の入力時のサイズは 512×512 画素であるが、実際に合成した画像は、これを圧縮しローパスフィルタをかけること

Fig. 5 System configuration

Fig. 6 Structure of 16 VITs

によりスムージングした、240×256 画素の画像である。この画像サイズは、2 MB プログラムメモリ中のプログラムのサイズによるが、512×512 画素程度まで任意に変更が可能である。

各 VIT はホストコンピュータからプログラムをダウンロードされた後、1~16 番号がふられる。このうち 1~14 番の VIT は描画用としてループ状に構成される。また 15, 16 番の VIT は入力画像の認識用に用いられ、認識結果は通常の通信リンクにより描画用 VIT に転送される。Fig. 6 に 16 台の VIT の構成を示す。これら 2 台の認識用 VIT の負荷分散は、等価な処理の時分割による。また画像の描画は、同様な等価な処理を描画用 VIT に時分割に実行させて行っている。認識用 VIT から描画用 VIT への手形状識別結果の受け渡しは、4, 7, 11, 14 番の各描画用 VIT からのデータ要求に認識用 VIT が応じるという形式で行われる。1~14 番の VIT には、人物画像のテキストチャデータが入力される。各 VIT により描画された画像は、1/30 秒ごとにモニタに転送される。

5.2 並列描画手法

通常、並列処理は負荷分散型の処理と機能分散型の処理に 2 分される。本システムではこのうち、主として以下の二つの理由により、人物モデルの描画を広義の負荷分散型の処理(異なる時刻の画像認識, 生成処理に異なるプロセッサを割り当てる時分割型並列処理)により行った。

- ① 負荷分散型処理によれば、一般に処理速度はプロセッサ数に比例させることができ、効率的にシステムの処理能力の向上を図ることができる。
- ② 一般に顔の表情の変化は、顔の各部分の協調性のある一連の動作と見ることができ、画像的には顔を形成する数多くのポリゴンが一定の関係を保ちつつ、規則正しく動くことを意味する。すなわち所定の表情に対し、移動すべき頂点とその方向、移動量などをあらかじめ定めることが可能である。このような場合、タスクの時間軸方向の分割、すなわち負荷分散型の処理が可能かつ有効である。

5.3 人物モデル

本研究で用いた人物モデルは、実際の人物像の上半身を用いた。このうちモデルの頭部においては、ワイヤフレームモデルを用いて表現している。このワイヤフレームモデルは、約 600 個の頂点よりなる約 500 個のポリゴン(微小 3 角平面)から構成されている。Photo. 2 に人物像頭部のワイヤフレームモデルをレンダリングした画像を示す。本画像においては、各ポリゴンの法線ベクトルの方向に基づき、可視部分かどうかの判別、および傾き角に応じた明度の設定処理を施している。

5.4 人物モデルの描画手法

5.4.1 基本手法

人物モデルの自然な 2 次元画像の合成には、3 次元モ

Photo. 2 Wire frame model

デリングとテクスチャ・マッピング法を適用している。テクスチャ・マッピングにおいては、高速化のために簡易化されたテクスチャ・マッピング法を考案し使用した。本手法は画素ごとに対応点を求めてマッピングするのではなく、対応するライン上のテクスチャデータを、データの補間を行いつつ一度にマッピングする方法であり、これによれば通常のレンダリングに要する時間とほぼ等しい時間で描画することができる。

5.4.2 人物モデルの表情合成

人物モデルでは、人物の顔の各部分(左右の瞼・左右の瞳・上下唇および顎)をつつまたは複数選択して反応させることができる。これにより、人物モデルに任意の表情をさせることができる。表情の設定には下記の各パラメータを要するが、本システムでは②～③の各パラメータはあらかじめ設定した値を用い、ユーザからの入力コマンドは①の部分の指定のみを行っている。

① 描画・表示部分：左瞼・右瞼・左右の瞳・上下唇および顎(口の周辺領域)から、描画・表示領域をつつまたは複数設定する。

② 頂点の移動距離：合成画像の頂点の初期位置からの移動距離を設定する。ここで設定した値は、それぞれ顔の各部分において挙動の有無を示し得る程度のもとした。また設定する移動距離は、代表点の移動距離であり、関連するほかの頂点の移動距離は、その全てに対しあらかじめ代表点の移動距離との関数関係を設定しているため、自動的に変換して算出し、設定される。たとえば Fig. 7 に示す目に関する頂点の移動においては、上瞼や瞳の代表点の移動と、関連するほかの頂点の移動の関係は、唯一に設定できる。

③ 描画ステップ：頂点の初期位置から移動位置までの、描画ステップ数を設定する。本システムでは、描画ステップを 3 に設定している。たとえば瞼は 3 ステップ(描画時の 3 フレーム)で完全に閉じる(開く)ことになる。

5.4.3 手の形状識別によるコマンド入力

手の形状識別による指の数と、顔の反応部位との対応

Fig. 7 Distance of movement of representative points

Fig. 8 Definition of coordinate

を以下に示す。ここでコマンド入力のある場合は、どのコマンドにおいても人物モデルは手の位置をトラッキングするものとする。また、瞳は顔と連動し、手を“見つめる”動作をする。

- 1本：顔の特定部位の指定はない。
- 2本：右目でウインクをする。
- 3本：瞬きをさせる。
- 4本：口をあける。
- 5本：すべての部位を動かす。

5.4.4 頂点の座標値の算出と座標系の設定

本システムでは、人物像頭部の首振り動作および表情の合成のため、Fig. 8に示すように系全体を記述する絶対座標系と、頭部内部における相対座標系の二つの座標系を設定した。頂点の座標値の算出式およびアルゴリズムを以下に示す。

(1) まず、相対座標系内において入力コマンドに依り、顔の各部分の頂点座標値の算出を行う。各 VIT は頂点の最大移動距離および描画ステップ数に基づき、逐次描画する頂点の位置を線形補間により算出している。たとえば(1)式において、 V_i, V_f, K, k をそれぞれ初期値、頂点の移動距離、描画ステップ、描画ステップ数とすると、頂点の描画位置 V_o は以下のようにして計算される。

$$V_o = V_i + (k/K) \cdot V_f \quad (1)$$

これによれば、描画ステップ数 k を適宜定めることにより、大きく開いた口や、少し開いた口などを描画し、表示することができる。

(2) (1)で求められた頭全体の座標値を、入力された手の位置を基に絶対座標系に変換する。描画は絶対座標系での頂点の座標値に対して行う。

5.4.5 描画および表示のタイミング

各 VIT はフレームメモリを一つずつ有するため、描画中に画像を表示することができない。すなわち描画中に表示すれば、描画過程そのものが表示されてしまう。そこで、描画用の 14 台のうち 1 台の VIT を順次画像の表示用に割り当てている。具体的には、画像表示中の VIT はそれ自身のもつ画像の表示のみを行い、つぎの VIT が描画を終えるとつぎの VIT に対し画像を表示す

るように命令する。表示を終えた VIT はつぎの表示のために描画を始める。この過程を順次繰り返す。各 VIT における画像の表示時間は、描画スピードに合わせ、任意に設定することができる。これは毎秒の表示コマ数の設定を意味する。

5.5 画像合成アルゴリズム

本節では、ホストコンピュータ(以下ホストとする)から TN-VIT にプログラムをダウンロードしてから画像合成に至るまでの、システムのアルゴリズムを示す。

- ① 各 VIT に 1~16 の番号を割付け、VIT の構成を Fig. 6 に示す構成とする。
- ② ホストより 1~14 番の VIT に人物画像のテクスチャのデータを転送する。
- ③ ホストより各 VIT に頂点の初期位置およびポリゴンのデータを転送する。
- ④ 1~14 番の各 VIT がバックグラウンド(背景)を描画する。その際、口の中のデータ(歯を含む)は原画像にないので、レンダリングにより背景に書き込む。各 VIT は、処理の間を通じてバックグラウンドを表示し続けるが、毎回の描画は処理時間の短縮のために、頭部の描画対象領域のみに対して行う。
- ⑤ 処理開始のコマンドをホストより TN-VIT に送る。
- ⑥ 画像認識用の VIT が入力画像中の手の位置と提示された指の数を計測し、その結果を描画用の VIT に転送、描画用の VIT は人物モデルの頭部を描画し、これをバックグラウンドと合成した画像を表示する。
- ⑦ 以後 1~14 番の VIT は、画像認識用 VIT より解析結果を入力しながら順次描画を行う。
- ⑧ ホストからの終了コマンドにより処理を終了する。

6. 実験および結果

実験は TN-VIT を中心とする、Fig. 5 に示したシステムにより行った。プログラムはトランスピュータ上での並列アルゴリズムの表現に適した言語、OCCAM により書かれている。ここでカメラに提示する手の位置は、カメラから 90~120 cm の距離で、カメラの光軸に対し上下に ± 20 cm、左右に ± 30 cm の範囲にあるものとする。また指と指の間は先端部で 1 cm 程度以上離れているものとする。

6.1 入力画像の認識精度および認識速度

まず、入力画像の認識精度および認識速度を調べる実験を行った。実験はカメラの視野中に各 100 回ずつ手形状を見せ、その正答数に基づき算出した。

Table 1 Hand shape recognition result

指の数	単位 (%)		
	条件	半袖	長袖
1 本		98	81
2 本		97	80
3 本		95	77
4 本		94	74
5 本		93	72

これによれば、半袖の場合、掌か手の甲のなす面の法線方向がカメラの光軸の方向とほぼ一致していれば、手の回転角に影響を受けることなく、各指の本数に対し平均 95% 以上の認識率が得られた。

しかし長袖の場合、服のしわなどにより認識率は平均 76.8% と低下した。また肌色やこれに近い配色や模様を

含む服に対しては、認識率は平均 50.6% であった。これは、袖のしわや模様が指先と誤認識されるためであった。

Table 1 に実験結果をまとめる。また Photo. 3~Photo. 6 に、手形状の識別処理の成功例と失敗例を示す。Photo. 3 は半袖の場合の 5 本指の識別処理の原画像であり、Photo. 4 はその処理結果である。また Photo. 5 は肌色・長袖の場合の 5 本指の識別処理の原画像であり、Photo. 6 はその識別処理が失敗した場合の例である。Photo. 6 中、画面右下の袖部分の処理結果が処理の失敗例である。しわと影の形状により指先と誤認識している。

また認識速度は、認識処理を 100 回行い、要した時間の平均を取るにより算出した。その結果、認識に要する時間は、約 0.5 秒であった。したがって入力画像の認識用に 2 台の VIT を割り当てた本システムでは、入力画像のサンプリングサイクルは、約 4 Hz であった。

Photo. 3 Original image for hand shape recognition 1
(Without sleeve, 5 Fingers)

Photo. 5 Original image for hand shape recognition 2
(With skin color sleeve, 1 Finger)

Photo. 4 Recognition result 1 (Success)

Photo. 6 Recognition result 2 (Failed)

Photo. 7 Original image of an Agent

Photo. 9 Synthesized image 2 (A nod of the agent)

Photo. 8 Synthesized image 1 (Movement of eyes and mouth)

6.2 TN-VIT の描画速度

1台の VIT は、二つのポリゴンにより 400×400 画素の画像に対してテクスチャマッピングする場合、約 3.7 秒で描画する性能を有する。また実験の結果、VIT 14 台を用いた時分割処理により、単に人物モデルを順次描画する場合、毎秒約 10 コマの描画が可能であった。またホストコンピュータと VIT システムとの各種コマンドの通信に要する時間は、約 10 ms であった。また本システムでは、人物モデルによる画像入力対象物の認識およびトラッキングと表情の合成速度は、毎秒最大約 8 コマであった。Photo. 7 にエージェントの原画像を示し、Photo. 8, Photo. 9 に出力画像例を示す。Photo. 8 は人物モデルの両眼を閉じさせ、口をあけさせた例である。また Photo. 9 は、人物モデルに頷きの動作をさせた例である。

6.3 考 察

実験の結果、画像入力された対象物の認識、および人物モデルによる表情の合成を含む対象物のトラッキングが、毎秒約 8 コマで実行可能であることが判明した。この処理速度は、描画処理では時分割による負荷分散処理

を行っているので、VIT の数を増やすことによりさらに向上させることが可能である。また本システムでは、手形状をカラー画像入力してその読み取りを行い、成果を得た。しかし半袖の場合は十分実用レベルの認識結果が得られたが、特に肌色やこれに近い長袖着用の場合に問題を残した。この問題の解決は、今後の重要な課題の一つである。今後は高度なヒューマンインタフェースの実現のために、人物表情の読み取りなど、より高度な画像認識および人物モデルの描画を、実時間で行う必要がある。また本稿では、時分割型の並列処理による画像の高速処理・描画に重点をおいたが、今後は人工感性を含めた高度な表情の合成のため、時分割型以外の並列処理手法の検討も必要であろう。

7. む す び

「ビジュアル・ソフトウェアエージェント、VSA」の実現のために、実際の人物像をテクスチャマッピングした人物モデルが、実時間でユーザの挙動を認識し反応する、画像認識・画像合成システム“TN-VIT”およびそのソフトウェア・アルゴリズムについて述べた。TN-VIT においては、32 bit 並列高速画像データベースを有する独自のトランスピュータボード VIT を開発し、これを 16 台製作して利用した。本稿で述べたシステム構成およびアルゴリズムにより、人物モデルが入力に応じ、表情を変えながら入力目標をトラッキングする姿を最大約 8 コマ/秒でカラー描画可能であることを確認した。本システムは、画像処理と画像合成を同一のシステム上で実時間で行うことを実現した、現在他に例をみないシステムであり、これからの画像システムの新たな方向を示すものと考ええる。今後はさらに VIT を増設し、より高速な画像処理・画像合成を行い、さらに通常のトランスピュータを標準通信リンクを介して並列接続し、より高度な画像認識や、知識ベースに基づく推論処理による人物モデル

のインテリジェント化を図る予定である。

本研究で使用した頭部ワイヤーフレームモデルは、本学工学部電子工学科原島(博)研究室より提供頂いたデータに基づき作成したことを記し、感謝します。またトランスピュータ開発環境の整備に御協力頂いた、SGS-THOMSON 社、松井和人氏に感謝します。

参 考 文 献

- 1) S. Brand 著, 室謙二, 麻生九美共訳: メディアラボ, 福武書店 (1988)
- 2) リチャード・ボルト著, 鶴岡雄二訳: マンマシン・インタフェース進化論, パーソナル・メディア (1990)
- 3) 石塚満: ヒューマンコミュニケーションと知識処理, 信学ヒューマンコミュニケーション研資, HC 90-5, (1990)
- 4) 郵政省電気通信局電気通信技術システム課(監修): ヒューマンインタフェース—人間中心主義のメディアステーションに向けて—, (財)日本データ通信協会 (1990)
- 5) 長谷川, W. Wongwarawipat, 李, 土肥, 石塚: Visual Software Agent による次世代ヒューマンインタフェース技術に関する研究, 電子情報通信学会秋期全国大会講演論文集, 6-257/258 (1990)
- 6) W. Wongwarawipat, C. W. Lee, O. Hasegawa, H. Dohi and M. Ishizuka: Visual Software Agent built on Transputer Network with Visual Interface, Transputing '91, 813/827, IOS Press (1991)
- 7) W. Wongwarawipat and M. Ishizuka: A Visual Interface for Transputer Network (VIT) and Its Application to Moving Image Analysis, Transputer/Occam Japan 3 (Proc. Third Transputer/Occam International Conference), IOS Press (1990)
- 8) 森島, 小林, 原島: マルチプロセッサ構成による知的画像符号化のためにリアルタイム表情合成の試み, 信学論(D-II), 73-D 2, 1647/1654 (1990)
- 9) INMOS Ltd.: Transputer Reference Manual, Prentice Hall (1988)
- 10) R A. Bolt: Put-That-There: Voice and Gesture at the Graphics Interface, ACM-SIGGRAPH, 14-3, 262/270 (1980)
- 11) 加藤, 吹野, 小山, 三輪: 立体図形編集における手振りインタフェース, 電子情報通信学会秋期全国大会講演論文集, 1-128 (1991)
- 12) 青島, 羽成, 吉田, 藪田: ニューラル型手話翻訳装置による指文字の認識, 電子情報通信学会秋期全国大会講演論文集, 1-129 (1991)
- 13) 石淵, 竹村, 岸野: 画像処理を用いた実時間手形状認識とマンマシンインタフェースへの応用, 電子情報通信学会秋

期全国大会講演論文集, 1-132 (1991)

- 14) 福本, 間瀬, 末永: 画像処理を用いた指示動作検出の実験システム, 電子情報通信学会春期全国大会講演論文集, 1-251 (1991)

[著 者 紹 介]

長谷川 修 (学生会員)

1988 年, 東京理科大学理工学部機械工学科卒業。90 年同大学院修士課程修了。同年より東京大学大学院電子工学専攻博士課程入学, 現在に至る。電子情報通信学会, 日本ロボット学会, 日本機械学会などの会員。画像理解, CG の研究に従事。

李 七 雨

1986 年, 韓国中央大学校工学部電子工学科卒業。88 年同大学大学院修士課程修了。92 年東京大学大学院博士課程修了。同年より(財)イメージ情報科学研究所勤務。画像理解, CG の研究に従事 (工学博士)。

Wiwat Wongwarawipat

1986 年, 電気通信大学工学部電子工学科卒業。88 年同大学大学院修士課程修了。91 年東京大学大学院博士課程修了。同年より(株)イーゼルロボットビジョン勤務。画像処理, 画像理解, 各種計算機の開発に従事 (工学博士)。

石 塚 満

1971 年, 東京大学工学部電子工学科卒業。76 年同大学大学院博士課程修了。同年 NTT 入社, 横須賀研究所勤務。78 年東京大学生産技術研究所助教授, 92 年同教授。人工知能, ビジュアルコンピューティングの研究を行っている (工学博士)。