TR-S-0019

3次元顔モデルを用いたビデオ映像中の自動顔トラッキングとモデルマッチムーブ

Automatic Face Tracking and Model Match Move by 3 Dimentional Facial Model

三澤貴文 Takafumi Misawa
村井和昌 Kazumasa Murai
中村哲 Toru Nakamura
森島繁生† Shigeo Morishima

2001.3.23

近年の著しい技術進歩により、携帯情報端末で動画像を送受信したり、音声翻訳システムを介して海外の人々と母国語で会話ができる時代もそう遠くはなく、音声のみならず画像も翻訳できれば会話がより自然なものになると思われる。この画像翻訳の実現には画像中の人物顔を正確にトラッキングする技術が必要となる。顔のトラッキングは多くの研究者により研究されてきたが、その多くが顔の特徴点を追うものであり、特徴点のフレーム間でのブレや顔の回転による特徴点の隠れなどの問題が残されていた。そこで、本論文では、画像翻訳の核となる技術として、3次元個人モデルを用いたテンプレートマッチングによる顔のトラッキング手法を提案した。そして、評価実験により、ある軸での角度の平均誤差が約0.28度という結果を得た。この結果は、提案した手法が効果的な方法であることを示すものである。

† 成蹊大学

1. はじめに

現在、テレビや映画など様々なメディアにおいてCG(Computer Graphics)が用いられており、そのクオリティはここ数年の間に飛躍的に進歩し、実写に近いレベルにまで達している。また人間の顔表情をCGにより実写同様にリアルに表現する研究が盛んに行われており、様々な分野での応用が期待されている。

通常、人間同士のコミュニケーションにおいては言葉による(バーバル)情報が最も重要であるが、それに加えて非言語的(ノンバーバル)な情報もまた非常に重要な情報源である。このノンバーバルな情報としては身振り・手振りといったものが考えられるが、特に多くの情報を発信しているのが顔の表情であることは容易に理解でき、従来よりコンピュータ上での顔モデルを介したコミュニケーションの研究が行われている。

また音声翻訳の研究においても、ATR音声言語通信研究所が発表した、話題の対象を限定するなどの一定の条件下で異なる言語間でのコミュニケーションを支援するツールATR-MATRIX[1][2]に見られるように、あらゆる言語間で盛んに行われており、その発展は目覚しいものがある。

一方で従来より海外の映画やTVなどの音声の吹き替えが行われているが、音声と口の動きが同期していないという問題が長い間あった。また、近年のハードウェア・ソフトウェア両面の技術進歩によって、携帯情報端末で当たり前のように動画像を送受信したり、音声翻訳システムを介して海外の人々と母国語を使用して会話ができる時代もそう遠くはなく、音声のみならず画像も翻訳できれば会話がより自然なものになるように思われる。

そこで本稿では、正確な3次元形状を持った顔モデルの生成から、ビデオ映像中の人物顔の移動や回転といった動きをテンプレートマッチングを用いて自動トラッキングし、その対象人物の顔を様々な表情・口形に合成可能な3次元モデルに置き換える画像翻訳システムについて述べる。

2. 3次元モデルの生成

人物の顔表情や口形状を合成する際に必要となる3次元個人モデルを取得する方法としては、対象人物の顔形状を測定して厳密なモデルを作り上げる方法と、ある標準的なモデルを用意して、それを対象人物に整合して使用する二通りの方法が考えられる。前者の方法は個人の正確な3次元形状を持つモデルが得られるものの、どこが目でどこが口であるかといった顔の各部位とモデルとの対応が取れていないという問題があり、また後者の方法は顔の各部位とモデルとの対応が取れているものの、個人の正確な3次元形状を再現するのが非常に難しいという問題がある。後者の方法による人物の顔画像を用いて個人の3次元顔モデルを生成するツールとしては、IPAから発表されている整合ツールなど[3][4]が知られている。しかしながら上述したように、人物の正面や側面といった限られた画像だけを使用したこれらのツールによる整合手法では、対象人物の正確な頭部形状を再現することは不可能である。

そこで本章では、Cyberware®[5]を用いた正確な3次元形状を持った個人顔モデルの生成手法について述べる。

2.1. 標準3次元頭部モデル

人間の顔は、基本的な形状や構造は同じと言ってよいが、目・鼻・口などの顔を構成する部位の形状や位置関係は個人個人によって微妙に異なる。そこで個人モデル作成の基となる標準モデルを用意した。図2-1に本稿で用いる標準3次元頭部ワイヤーフレームモデルを示す。このモデルは頂点数が759点で、1352個のポリゴンにより構成されている。

2.2. Cyberware®を用いた3次元顔モデル生成

本稿で紹介するテンプレートマッチングを用いた顔の自動トラッキングを精度よく行うには、個人の顔形状に忠実な顔モデルを作成する必要がある。そこで本節では、目や口などの点の対応が取れた標準ワイヤーフレームモデルに、Cyberware®(図2-2)を用いて取得した人物の顔の正確な3次元形状を当てはめることで、正確な3次元形状を持っ

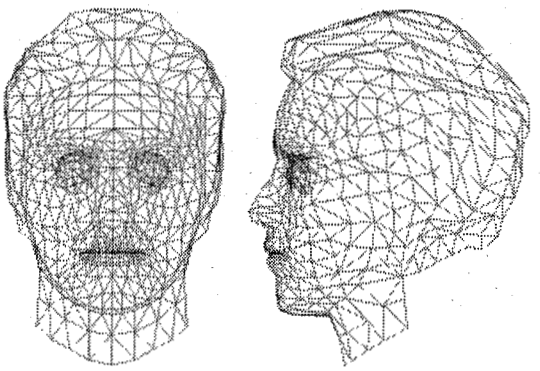

図2-1：標準頭部3次元ワイヤーフレームモデル

図2-2：Cyberware®

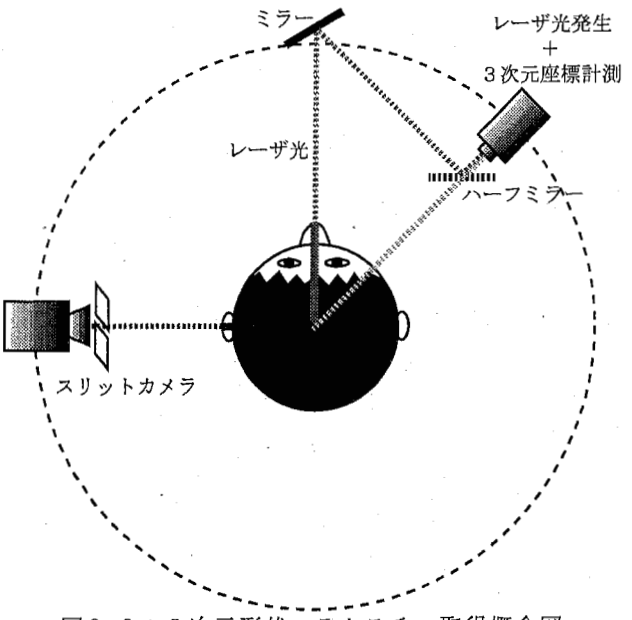

図2-3：3次元形状・テクスチャ取得概念図

図2-4：取得された3次元形状とテクスチャ

た個人の顔モデルを生成する手法について述べる。

2.2.1. Cyberware® の動作原理

Cyberware® とは頭部の周りをカメラが一回転して、頭部の3次元座標とテクスチャを取得する装置である。具体的には、1周360°を512に分け、360/512°回転する毎に赤いレーザ光を頭部に当てて、原点から赤いレーザ光の当たっている輪郭までの距離を測定することにより頭部の3次元座標を取得する。テクスチャに関しては回転中に連続的に頭部周囲のテクスチャをスリットカメラにより撮影する。3次元形状・テクスチャ取得の概念図を図2-3に、Cyberware® を使用して得られた3次元座標及びテクスチャを図2-4に示す。

2.2.2. 標準モデルとの整合

Cyberware® により取得した頭部の3次元モデル座標（以下、Cyberware モデル）と標準ワイヤーフレームモデルを整合する。整合は、標準ワイヤーフレームモデル及びCyberware モデル共に2次元に展開後、各々の顔の各部位を2次元平面上でマニュアル整合し、最後に標準モデルの座標を対応するCyberware モデルの座標に置き換えることで対象人物に忠実な3次元形状を持った顔モデルを生成した。図2-5に整合前後の様子、図2-6に生成した顔モデルを示す。

3. 口形動作の規則化

人物の顔には喜怒哀楽による表情変化や、発話することによる口形の変化があり、会話の理解においてこのような唇の動きが特に重要な役割を果たしていることも研究[6]により分かっている。その為、3次元モデルを変形して顔画像を合成するには、その発話時の人間の口領域を忠実に再現する必要がある。また、人間が会話している際に変化量が大きい部分としては、唇・顎などが挙げられるが、特に唇の動きは音韻と密接な関係があるため、口形状の変化を表現するための特徴点を細かく設定する必要がある。

口領域の動きを定量的に表現するために、口領域の制御点として図3-1に示す①～⑪までの11点を定めた。①は上唇の上端、②は上唇の下端、③は下唇の上端、④は下唇の下端、⑤は顎の下端、⑥は唇の外端点、⑦は唇の内端点、⑧は上唇上側の① - ⑥及び① - ⑦の中点付近の格子点、⑨は上唇下端の② - ⑥及び② - ⑦の中点付近の格子点、⑩は下唇上端の③ - ⑥及び③ - ⑦の中点付近の格子点、そして⑪は下唇下端の④ - ⑥及び④ - ⑦の中点付近の格子点を示している。また、これらの制御点が動いたときに格子点が変形するための規則化を行う必要があるが、本稿では格子点の移動方法として、制御点①、②、⑦はx,y,z方向に移動し、③、④、⑤、⑥はx,y,z方向の移動に加えて回転の動きをさせている。図3-2に回転軸の位置を示す。以上のような動きを定量的に表現するため、表に示す17個のパラメータを定めた。ここで個人によって口の大きさが違うので、それが移動量に影響してくる。そのため横

図2-5：標準モデルとCyberware モデルの整合

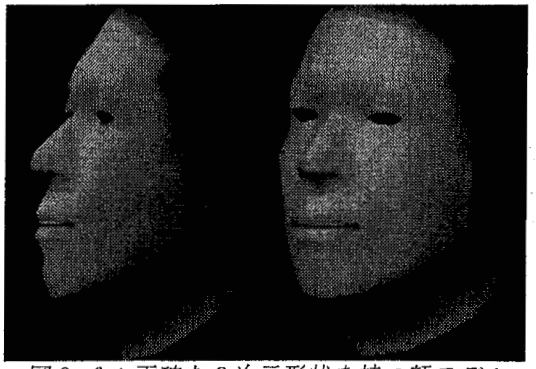

図2-6：正確な3次元形状を持つ顔モデル

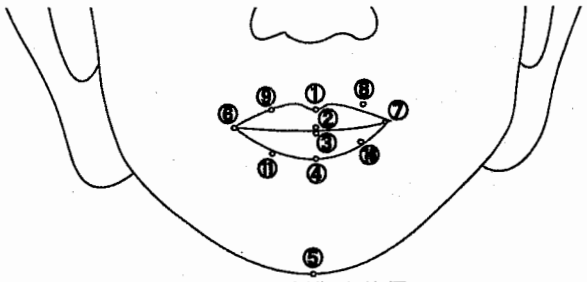

図3-1：制御点位置

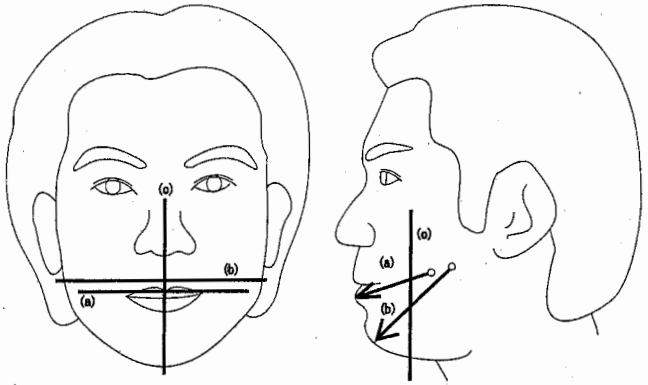

図3-2： 回転軸の位置

表：口形パラメータ

制御点	口形パラメータ	動作対象
1	1	上唇上側の縦方向の動きに対応するパラメータ
	2	上唇上側の奥行き方向の動きに対応するパラメータ
2	3	上唇下側の縦方向の動きに対応するパラメータ
	4	上唇下側の奥行き方向の動きに対応するパラメータ
3	5	下唇上側の縦方向の動きに対応するパラメータ
	6	下唇上側の奥行き方向の動きに対応するパラメータ
4	7	下唇下側の縦方向の動きに対応するパラメータ
	8	下唇下側の奥行き方向の動きに対応するパラメータ
5	9	顎の縦方向の動きに対応するパラメータ
6	10	唇端点の縦の動きに対応するパラメータ
	11	唇端点の横の動きに対応するパラメータ
	12	唇端点の奥行き方向の動きに対応するパラメータ
7	13	唇の横方向の開き具合に対応するパラメータ
8	14	上唇上側の制御点付近以外の動きに対応するパラメータ
9	15	上唇下側の制御点付近以外の動きに対応するパラメータ
10	16	下唇上側の制御点付近以外の動きに対応するパラメータ
11	17	下唇下側の制御点付近以外の動きに対応するパラメータ

方向の動きについては、唇の制御点同士の距離、すなわち制御点⑥と⑦の距離、 縦方向の動きについては鼻の頂点と顎の制御点の距離、すなわち図3-1の鼻の頂点と⑤の距離を基準にして正規化を行った。

4．3次元モデルを用いた自動顔トラッキング

ビデオ映像中の顔をトラッキングする方法としては、画像上の顔の特徴点を追うという方法[7 10]が長い間多くの研究者によって試みられてきた。しかしながらこのような1次元の点を追う方法には、認識された特徴点がフレーム間でブレたり、追跡している点が顔の回転等により隠れてしまって認識できないなどの問題があった。そこで本章では、特徴点という1次元の追跡ではなく、フレーム間でのブレや特徴点の隠れに影響されない3次元モデルを用いた2次元平面でのトラッキング手法を紹介する。

4．1． ビデオ映像中の人物動作に関する条件

今回使用したビデオ映像は、以下のような条件を付けて撮影した。

（1） 頭は普段の会話時に自然発生する程度の動き
（2） 探索の次元数を減らすために無表情
（3） 口は探索に使用しないので動いても可
（4） 顔に影などができないような光源の設置
（5） 顔の大きさが不変の（ズーミングが無い）もの

4．2． テンプレートの作成

ビデオフレーム画像とのテンプレートマッチングを行うため、2章で生成した3次元モデルにビデオ映像中の任意の1フレームのテクスチャを貼り、その3次元顔モデルにx，y方向への平行移動とx，y，z軸での回転を移動量・回転角を少しずつ変えながら適用して、その都度生成される2次元画像をテンプレートとして使用した。このときビデオ映像中の人物の口が動くことを考え、マッチング結果への影響を減らすためにテンプレートの口周辺部分を除いた。図4-1にテンプレート例を示す。

4．3． テンプレートマッチング

4．2．のテンプレートを作成する毎に、そのテンプレートのブルーバックを除いた部分（顔領域）とビデオフレーム画像とのテンプレートマッチングを行い、正規化誤差最小時の3次元顔モデルの位置及び角度を求め、その値をそのビデオフレームでの人物の顔の位置及び角度とした。ビデオ映像中の1フレームにおける顔の位置・角度を求める

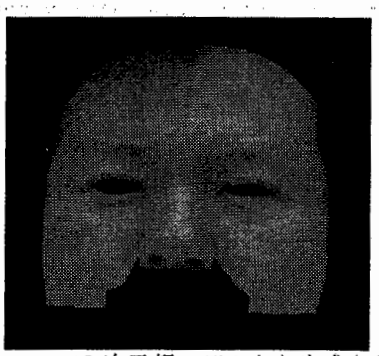

図4-1： 3次元顔モデルから生成される2次元画像（テンプレート）の例

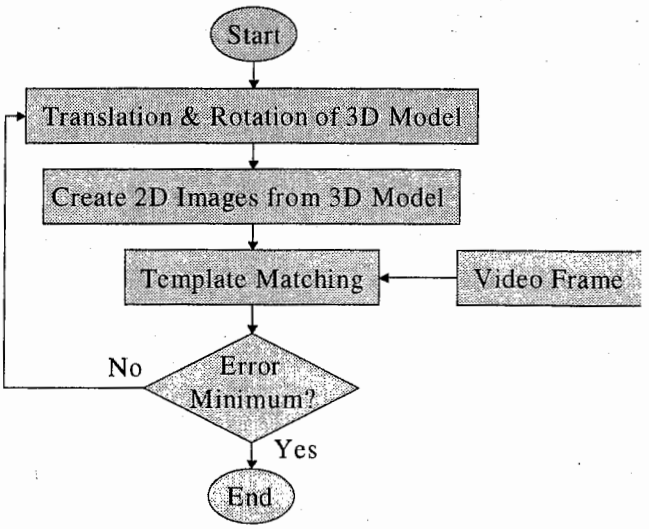

図4-2：最小誤差探索のフローチャート

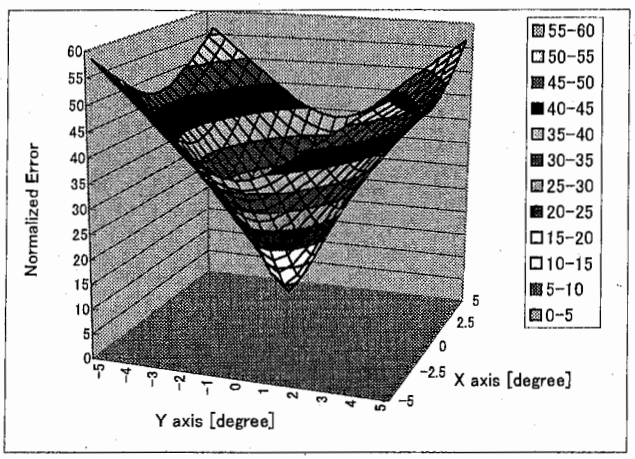

図4-3：x, y 軸を各々-5～5°まで1°づつ変化させていったときの正規化誤差

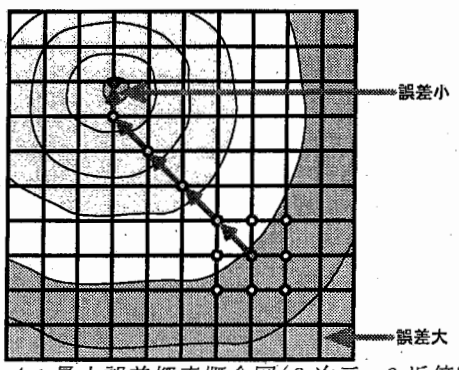

図4-4：最小誤差探索概念図(2次元、8近傍時)

フローチャートを図4-2に示す。ここで、誤差を求めるときにテンプレートのブルーバックを除いた顔部分のみでマッチングを行っていてテンプレート毎にマッチングするピクセル数が異なるので、式4-1, 4-2のように誤差を正規化してから誤差の比較を行っている。

$$Error = |R_V - R_T| + |G_V - G_T| + |B_V - B_T| \qquad (式4-1)$$

$$\text{Normalized Error} = \frac{Error}{\text{Number of Pixel}} \qquad (式4-2)$$

式4-1及び4-2の変数はそれぞれビデオフレーム画像のR,G,B値(R_V,G_V,B_V)及びテンプレート画像のR,G,B値(R_T,G_T,B_T)を表している。

4.4. 誤差最小値の探索

4.3.で述べたテンプレートマッチングによる探索は、必要な探索範囲全ての正規化誤差を求めた結果として答えが1つ決まるものである。ここで"必要な探索範囲"とは、4.1.の条件(1)を踏まえて撮影されたビデオ映像中の、フレーム間での最大移動量及び回転角を含む範囲のことであるが、この範囲全体の誤差を求めるのは非常に時間がかかる上に、不必要な値も含んでいると考えられるので無駄である。図4-3にx, y軸方向に各々-5°から5°まで1°づつ回転させた時の正規化誤差のグラフを示すが、このようなグラフより、探索範囲中に誤差のローカルミニマムは1つしか存在しないとの仮定をし、探索は前フレームでの位置・角度を起点として、起点の3^n-1近傍(nは次元数：5次元のとき242近傍)の正規化誤差を調べて誤差最小の点に順次移っていき、起点が誤差最小になったときの値をそのフレームでの位置・角度とするという方法で探索コストを抑えた。図4-4に誤差の最小値を探索する時の概念図(2次元・8近傍の場合)を示す。

5. OPTOTRAK® を用いた頭部の動き計測

第4章で3次元モデルを用いたビデオ映像中の顔トラッキングを行った。しかしながら、そこで得られた結果が数値的に本当に正しいのかは分からない。そこで本章では、3次元モデルを用いたトラッキング精度の評価の為、3次元運動計測システムOPTOTRAK®[11]を用いた頭部の運動計測について述べる。

5.1. OPTOTRAK® の動作原理

OPTOTRAK®は、人物に赤外線ダイオードを用いたマーカを取り付け、それを3台の赤外線CCDカメラを用いて計測するモーションキャプチャシステムである。本稿では人物頭部の移動量・回転角を測定するのでマーカを4つ使用した。また、3次元モデルを用いたトラッキングを顔部分で行うため、図5-1のようなマーカを頭部に取り付けた。図5-2にOPTOTRAK®による測定の概念図を示す。

5.2. 頭部の動き計測

頭部の動きとして、次のような動きを測定した。

(1) x軸方向の回転運動　(首を縦に振る運動)

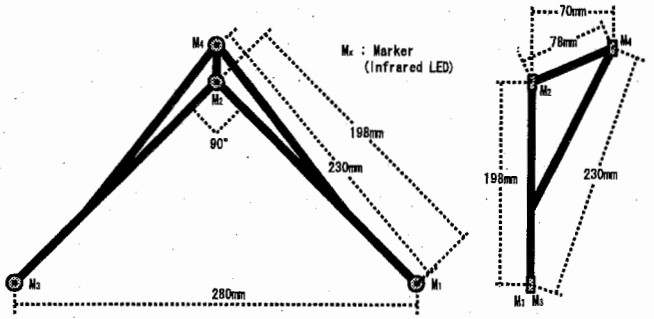

図5-1：頭部に付ける赤外線LEDマーカ

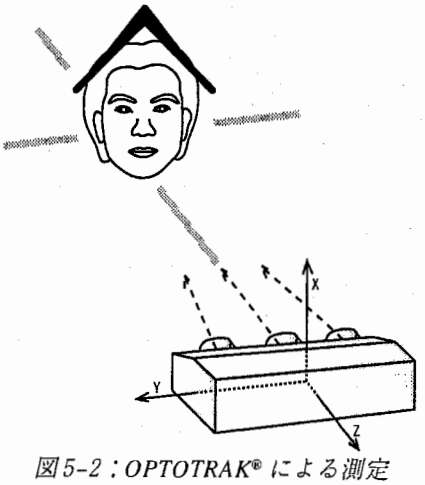

図5-2：OPTOTRAK® による測定

（2） y 軸方向の回転運動　（首を横に振る運動）
（3） z 軸方向の回転運動　（首を傾げる運動）
（4） 自由な動き
これらOPTOTRAK®による測定データをトラッキングの正解値として評価に用いた。

6． 画像翻訳システム

本章では、音声認識・翻訳・合成システムの紹介とそのシステムを使用して合成した音声を用いた画像翻訳システムについて述べる。

6.1． 音声認識・翻訳・合成

本画像翻訳システムの音声翻訳部分には、ATR 音声言語通信研究所から発表されている ATR-MATRIX を使用した。ATR-MATRIX とは音声認識システム（SPREC）・テキスト翻訳システム（TDMT）・音声合成システム（CHATR）の総称であり、各システムの概要は以下のようである。

6.1.1． 音声認識 （SPREC）

音声認識では、入力された音声を特徴ベクトルと呼ばれるスペクトル情報 x に変換し、音声データベースから構築した音響モデル $P(x|w_i)$ 及び言語モデル $P(w_i)$ から、条件付き確率 $P(w_i|x)$ が最も大きくなるような単語 w_i を探索して、認識されたテキストと単語の区切り情報を出力する（式 6-1, 6-2）。

$$P(w_i|x)=\frac{P(x|w_i)P(w_i)}{P(x)} \qquad \text{（式 6-1）}$$

$$k=\arg\max_i P(w_i|x) \qquad \text{（式 6-2）}$$

式 6-2 は、条件付き確率 $P(w_i|x)$ が最も大きくなる i を探して k に出力することを表している。このシステムは、不明瞭な発声でも認識できる音響モデルを提案しており、高精度の音声認識が可能となっている。

6.1.2． テキスト翻訳 （TDMT）

テキスト翻訳では、話し言葉特有の言い回しとその対訳文に基づいてパターンと用例をデータベース化し、これと入力文の類似性を調べながら、覚えている表現を使い適切に翻訳する。例えば「A はB へ行く。」という文章を” A goes to B.” という文にあらかじめ対応させてデータベースにしておき、「彼は駅に行く。」などのテキストが入力されてきたときに、データベースから最適な翻訳文を取り出すという方法で翻訳を実現している。

6.1.3． 音声合成 （CHATR）

音声合成では、TDMT の出力するテキスト及び単語の区切り情報を入力として、声の大きさ・高さ・長さを設定する。そして、それらの情報にマッチする適切な音声波形を音声データベースの中から選択して、必要により信号処理を行うことにより、極めて高品質な合成音声を生成するものである。このとき、音声認識で得た話者の発声の特徴パラメータも反映されるため、女性が喋れば女性の音声を合成する。

6.2． システム構成

図 6 に画像翻訳システムの全体の流れを示す。

音声は ATR-MATRIX により処理され、合成音声と口形状変形のための発音記号及び音素継続長に変換される。また、フレーム画像は第 2 章による 3 次元個人モデル作成を経て、第 4 章で述べたフレーム毎のマッチング処理が行われる。最後に ATR-MATRIX による発音記号・音素継続長情報により口形状を合成して画像翻訳が完了する。

ここでのトラッキング処理及び合成処理は以下のようなハードウェア構成で行った。

CPU : PentiumIII 1GHz x 2
Memory : 512 Mbytes
OS : Microsoft Windows 2000
Video card : Canopus SPECTRA 8800

なお、3 次元モデルによる合成画像をビデオフレーム画像に重ねる際、通常の画像の置き換えだけでは画像境界が出てしまうので、画像境界部分においては色を混色する処理をした。

7． システム評価

ここまで、Cyberware® データを用いた 3 次元モデルの生成から、そのモデルを使用したトラッキング、トラッキング精度評価のための OPTOTRAK® を用いた頭部の運動測定、そして画像翻訳システムについて述べてきた。

本章では、システム全体の評価について述べる。

7.1． Cyberware® フィッテングの評価

個人モデルの作成手法には様々なものがあるが、今回 Cyberware® により得られた 3 次元形状データを用いることによって、1 枚の画像との整合のみで正確な個人形状を持つ 3 次元モデルを作成できた。また、様々な角度から撮影された個人画像を n 枚使用して標準モデルを整合するような 3 次元モデル生成手法[4]と比べると、整合時間は約 1/n の時間で整合でき、形状も格段に良くなった。

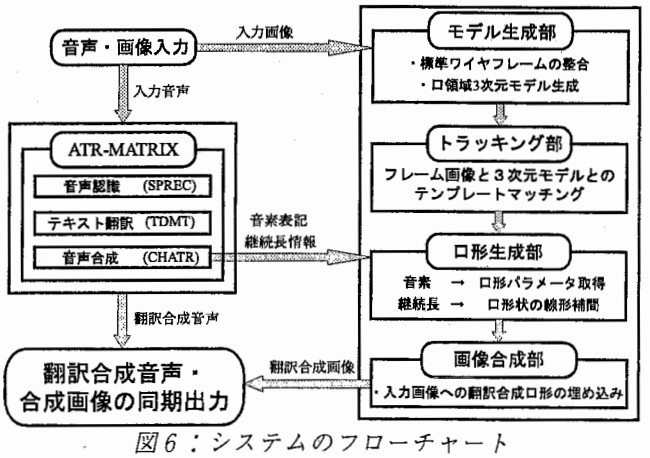

図6：システムのフローチャート

（2） y 軸方向の回転運動　（首を横に振る運動）
（3） z 軸方向の回転運動　（首を傾げる運動）
（4） 自由な動き
これらOPTOTRAK®による測定データをトラッキングの正解値として評価に用いた。

6. 画像翻訳システム

本章では、音声認識・翻訳・合成システムの紹介とそのシステムを使用して合成した音声を用いた画像翻訳システムについて述べる。

6.1. 音声認識・翻訳・合成

本画像翻訳システムの音声翻訳部分には、ATR 音声言語通信研究所から発表されている ATR-MATRIX を使用した。ATR-MATRIX とは音声認識システム（SPREC）・テキスト翻訳システム（TDMT）・音声合成システム（CHATR）の総称であり、各システムの概要は以下のようである。

6.1.1. 音声認識 （SPREC）

音声認識では、入力された音声を特徴ベクトルと呼ばれるスペクトル情報 x に変換し、音声データベースから構築した音響モデル $P(x|w_i)$ 及び言語モデル $P(w_i)$ から、条件付き確率 $P(w_i|x)$ が最も大きくなるような単語 w_i を探索して、認識されたテキストと単語の区切り情報を出力する（式 6-1, 6-2）。

$$P(w_i|x)=\frac{P(x|w_i)P(w_i)}{P(x)} \qquad \text{（式 6-1）}$$

$$k=\arg\max_i P(w_i|x) \qquad \text{（式 6-2）}$$

式 6-2 は、条件付き確率 $P(w_i|x)$ が最も大きくなる i を探して k に出力することを表している。このシステムは、不明瞭な発声でも認識できる音響モデルを提案しており、高精度の音声認識が可能となっている。

6.1.2. テキスト翻訳 （TDMT）

テキスト翻訳では、話し言葉特有の言い回しとその対訳文に基づいてパターンと用例をデータベース化し、これと入力文の類似性を調べながら、覚えている表現を使い適切に翻訳する。例えば「A はB へ行く。」という文章を”A goes to B.”という文にあらかじめ対応させてデータベースにしておき、「彼は駅に行く。」などのテキストが入力されてきたときに、データベースから最適な翻訳文を取り出すという方法で翻訳を実現している。

6.1.3. 音声合成 （CHATR）

音声合成では、TDMT の出力するテキスト及び単語の区切り情報を入力として、声の大きさ・高さ・長さを設定する。そして、それらの情報にマッチする適切な音声波形を音声データベースの中から選択して、必要により信号処理を行うことにより、極めて高品質な合成音声を生成するものである。このとき、音声認識で得た話者の発声の特徴パラメータも反映されるため、女性が喋れば女性の音声を合成する。

6.2. システム構成

図 6 に画像翻訳システムの全体の流れを示す。

音声は ATR-MATRIX により処理され、合成音声と口形状変形のための発音記号及び音素継続長に変換される。また、フレーム画像は第 2 章による 3 次元個人モデル作成を経て、第 4 章で述べたフレーム毎のマッチング処理が行われる。最後に ATR-MATRIX による発音記号・音素継続長情報により口形状を合成して画像翻訳が完了する。

ここでのトラッキング処理及び合成処理は以下のようなハードウェア構成で行った。

CPU : PentiumIII 1GHz x 2
Memory : 512 Mbytes
OS : Microsoft Windows 2000
Video card : Canopus SPECTRA 8800

なお、3 次元モデルによる合成画像をビデオフレーム画像に重ねる際、通常の画像の置き換えだけでは画像境界が出てしまうので、画像境界部分においては色を混色する処理をした。

7. システム評価

ここまで、Cyberware® データを用いた 3 次元モデルの生成から、そのモデルを使用したトラッキング、トラッキング精度評価のための OPTOTRAK® を用いた頭部の運動測定、そして画像翻訳システムについて述べてきた。

本章では、システム全体の評価について述べる。

7.1. Cyberware® フィッテングの評価

個人モデルの作成手法には様々なものがあるが、今回 Cyberware® により得られた 3 次元形状データを用いることによって、1 枚の画像との整合のみで正確な個人形状を持つ 3 次元モデルを作成できた。また、様々な角度から撮影された個人画像を n 枚使用して標準モデルを整合するような 3 次元モデル生成手法[4]と比べると、整合時間は約 1/n の時間で整合でき、形状も格段に良くなった。

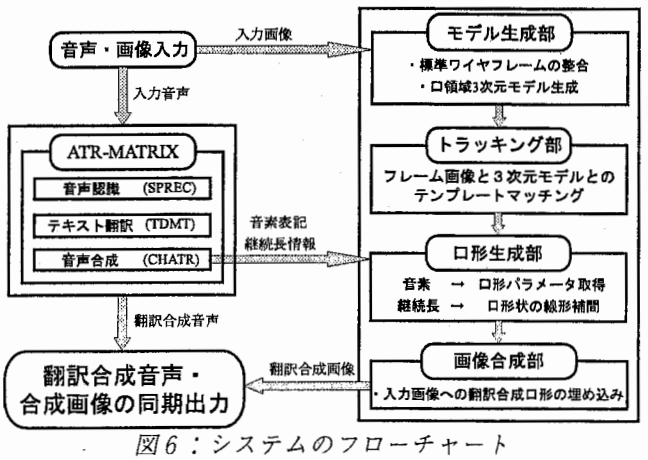

図 6：システムのフローチャート

7.2. トラッキング・モデルマッチムーブの評価

トラッキング結果として、各回転運動におけるテンプレートマッチングでのトラッキングとOPTOTRAK®でのトラッキングの角度比較のグラフを図7-1～3に示す。このグラフの角度0°は顔が正面を向いているときである。誤差の平均は図7-1のx軸が4.18°、図7-2のy軸が0.48°、図7-3のz軸が0.28°であった。

トラッキングの結果は撮影時の光源の設置場所に大きく影響される。図7-2,3を見るとトラッキング結果の誤差は少ないが、図7-1の結果は正解値から大きくズレている。前者の誤差は、テンプレートマッチングに使用する3次元モデルに、正面画像から取ったテクスチャをマッピングしており、正面方向以外から見たときの画素情報が不正確であるために主に生じたものであると考えられる。しかしながら後者（前者にも言える）の誤差は、今回のOPTOTRAK®での評価用ビデオ撮影時の光源配置に問題があり、x軸上の回転、つまりうなづいたときに顔全体に影ができ、ビデオフレーム画像の顔部分の方がテンプレート画像の顔より色が黒くなってしまったために、トラッキング結果に影響したものと思われる。

トラッキングの処理時間は、現在ビデオカードによる影響をかなり受けるが、今回の探索アルゴリズム、システム構成、ビデオ映像を用いた結果では1フレーム平均約28.69[s]であった。これは頭部の動く速度などによってかなり変動する。今回ハードウェア構成としてデュアルプロセッサのものを使用しているが、今回の探索ではシングルプロセッサで動作させている。

トラッキング結果を使用してビデオフレーム画像中の顔を3次元モデルに置き換えるモデルマッチムーブ処理をした。モデルマッチムーブの様子を図7-4に示す。ここで顔部分だけを置き換えるだけでは3次元モデルとビデオフレーム画像との境界がはっきりと出てしまい不自然なものとなるので、境界部分で色の混色を行っている。このことと高精度のトラッキング結果によって、画像境界及び顔領域を3次元モデルに置き換えていることすら分からないような結果を得た。

7.3. 翻訳システムの評価

6章で述べたシステムにより翻訳音声と合成口形を同時出力した。画像翻訳をしなかった時と比べると幾分自然になっているように思われるが、今回はこの点についての評価をしていない。そこで今後、翻訳音声と画像翻訳無し、翻訳音声と画像翻訳有りなどといった様々な音声と画像の組み合わせによる主観評価実験を行い、翻訳システムの評価を行う予定である。

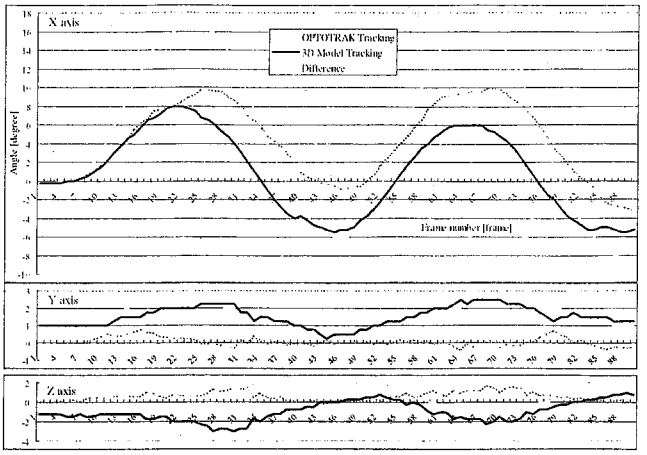

図7-1：主にx軸上の回転運動における各軸のトラッキング結果と誤差（上からx軸、y軸、z軸）

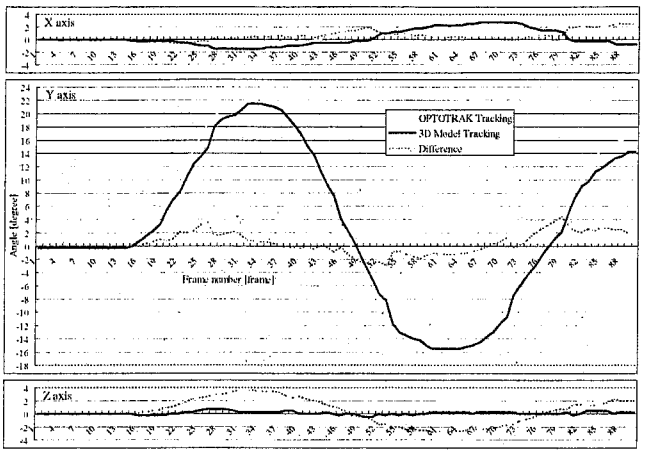

図7-2：主にy軸上の回転運動における各軸のトラッキング結果と誤差（上からx軸、y軸、z軸）

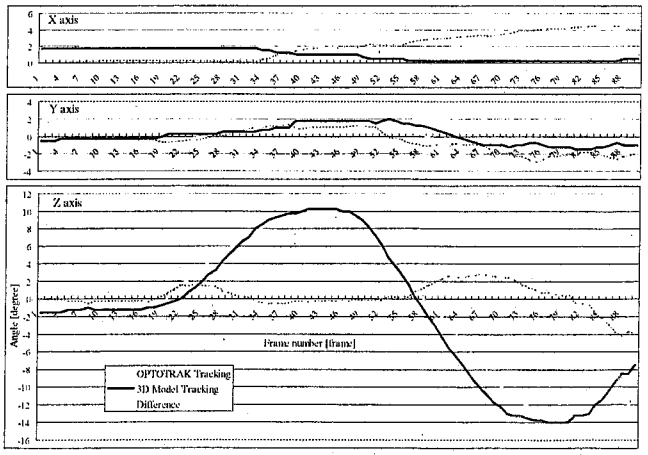

図7-3：主にz軸上の回転運動における各軸のトラッキング結果と誤差（上からx軸、y軸、z軸）

8. まとめ

人物の顔形状に、より忠実な3次元個人モデルの作成として、Cyberware®により取得した3次元形状データを用いた標準モデルへの整合を行うことで、今まで多大な時間を要していた整合時間が大幅に短縮され、かつ、対象人物に忠実な3次元形状を持った個人モデルを生成できるようになった。

人物頭部の動きのトラッキングに関しては、顔

図7-4：モデルマッチムーブ（上段：原画像、中段：重ねる顔領域、下段：合成画像）

の特定の部位のテンプレートマッチングによる特徴点の追跡ではなく、3次元モデルを移動・回転させてできる顔画像全体を用いたテンプレートマッチングを行って移動量・回転角を求めるという方法によって、今までの多くのトラッキングアルゴリズムにおいて問題となっていた特徴点のブレや回転による隠れの影響を受けない、非常に精度の高いトラッキング結果を得ることができた。

トラッキング結果を使用して3次元モデルをビデオ画像に重ね合わせるモデルマッチムーブ処理、口形状を変化させる画像合成に関しては、ビデオ画像への重ね合わせの際に、3次元モデルの周囲テクスチャとの混色を行うことにより、一見、3次元モデルが重なっているとは分からないような効果を得た。

今回のトラッキングでは、無表情でカメラのズーミングのないものを主に取り扱った。しかしながら、実際の人物にはこのような動きが必ず含まれており、表情やズーミングといった動きを入れたときにトラッキング結果への影響が少なからず発生すると考えられる。今後はそれらの影響の測定や対応が課題である。これらが解決されれば、3次元モデルを使用したテンプレートマッチングによる表情認識や口形推定なども行えるようになると思われる。

また、今回の最小値探索アルゴリズムを見直すことにより、更なるトラッキングの高速化を計ることが出来ると思われる。トラッキングをリアルタイムで行えるようなアルゴリズムを考え、撮影された映像を即座に3次元モデルに置き換えて合成できるようにすることも今後の課題である。

参考文献

[1] 菅谷 史昭、竹澤 寿幸、横尾 昭男、山本 誠一「日英双方向音声翻訳システム(ATR-MATRIX)の対話実験」日本音響学会1999年春季研究発表会講演論文集 pp107-108, 1999

[2] ATR-MATRIX Home Page : http://www.itl.atr.co.jp/matrix/cstar/

[3] 「独創的情報技術育成事業に係わる感性擬人化エージェントのための顔情報処理システムの開発最終成果報告書」財団法人 イメージ情報科学研究所

[4] 伊藤 圭、三澤 貴文、武藤 淳一、森島 繁生「複数アングル画像からの3次元頭部モデルの生成と表情合成」電子情報通信学会技術研究報告 Vol.99, No.582, pp7-12, 2000

[5] Cyberware Head & Face Color 3D Scanner Web Site : http://www.cyberware.com/products/index.html

[6] Neely, K.K. "Effect of visual factor on the intelligibility of speech", J. Acoustic. Soc. Amer., Vol.28, p1275, 1956

[7] 永田 亮一、川口 剛「複雑な背景を持つカラー画像からの顔検出」電子情報通信学会技術研究報告 PRMU' 99-111, pp.77-82, 1999

[8] 吉村 哲也、市川 忠嗣、森島 繁生、相澤 清晴、斎藤 隆弘「空間コミュニケーションにおける表情入力のための顔抽出」1999年映像情報メディア学会冬季大会, P.35, 1999

[9] 島田 直幸、武藤 淳一、森島 繁生「Pan Tilt Zoom Controllable Cameraによる目および唇形状抽出・追跡」電子情報通信学会 1999年総合大会講演論文集, D-12-71, p.244, 1999.3

[10] 四倉 達夫、島田 直幸、森島 繁生、大谷 淳「アクティブカメラによる視線追跡・自動Lip Reading」電子情報通信学会技術研究報告 Vol.HIP99-38, No.452, pp.31-36, 1999.11

[11] 3次元運動計測システムOPTOTRAK Web site : http://www.asco.co.jp/asco/products/optotrak.html