

(19) 日本国特許庁 (JP)

(12) 公 開 特 許 公 報 (A)

(11) 特許出願公開番号

特開2013-149251

(P2013-149251A)

(43) 公開日 平成25年8月1日 (2013. 8. 1)

(51) Int.Cl.	F I	テーマコード (参考)
G 0 6 T 13/40 (2011.01)	G O 6 T 13/40	5 B O 5 O
G 1 0 L 15/00 (2013.01)	G 1 0 L 15/00 2 O O Z	

審査請求 有 請求項の数 13 O L 外国語出願 (全 40 頁)

(21) 出願番号	特願2013-6090 (P2013-6090)	(71) 出願人	392026693
(22) 出願日	平成25年1月17日 (2013. 1. 17)		株式会社エヌ・ティ・ティ・ドコモ
(31) 優先権主張番号	12151415.2		東京都千代田区永田町二丁目11番1号
(32) 優先日	平成24年1月17日 (2012. 1. 17)	(74) 代理人	100099623
(33) 優先権主張国	欧州特許庁 (EP)		弁理士 奥山 尚一
		(74) 代理人	100096769
			弁理士 有原 幸一
		(74) 代理人	100107319
			弁理士 松島 鉄男
		(74) 代理人	100114591
			弁理士 河村 英文
		(74) 代理人	100125380
			弁理士 中村 綾子
		(74) 代理人	100142996
			弁理士 森本 聡二

最終頁に続く

(54) 【発明の名称】 顔の口をアニメーション化するコンピュータ実施方法及び装置

(57) 【要約】 (修正有)

【課題】 所与の視覚音素シーケンスに従って顔の口をアニメーション化するコンピュータ実施方法を提供する。

【解決手段】 フェイスモデルにより顔をグラフィックで表現するステップと、複数のV個の可能な視覚音素の各々に、その視覚音素の複数のI個の異なるスキャン又はサンプルを測定により取得するステップと、上記フェイスアニメーションモデルにより上記複数のV個の異なる視覚音素の各々の上記複数のI個の各サンプルを表現して、視覚音素シーケンスを視覚音素空間を通る軌跡を通じて表現することができるように、上記視覚音素空間に広がる上記スキャン又はサンプルに基づいて行列を生成する、表現するステップと、ベイズ手法を適用して、上記所与の視覚音素シーケンスの上記視覚音素空間を通る最良のパスを取得する、適用するステップと、を含む、所与の視覚音素シーケンスに従って顔の口をアニメーション化するコンピュータ実施方法。

【選択図】 なし

【特許請求の範囲】

【請求項 1】

所与の視覚音素シーケンスに従って顔の口をアニメーション化するコンピュータ実施方法であって、

フェイスモデルにより前記顔をグラフィックで表現するステップと、

複数の V 個の可能な各視覚音素に、該視覚音素の複数の I 個の異なるスキャン又はサンプルを測定により取得するステップと、

前記フェイスアニメーションモデルにより前記複数の V 個の異なる各視覚音素の前記複数の I 個の各サンプルを表現して、視覚音素シーケンスを視覚音素空間を通る軌跡を通じて表現することができるように、前記視覚音素空間に広がる前記スキャン又はサンプルに基づいて行列を生成する、表現するステップと、

ベイズ手法を適用して、前記所与の視覚音素シーケンスの前記視覚音素空間を通る最良のパスを取得する、適用するステップと、
を含み、

前記視覚音素リストを所与とした前記軌跡の確率

【数 1】

$$p(\beta|\mathcal{L}) \sim p(\mathcal{L}|\beta) \cdot p(\beta)$$

は、前記軌跡を所与とした前記視覚音素リストの確率

【数 2】

$$p(\mathcal{L}|\beta) = \prod p(v_i|\beta(t_i))$$

に比例する、所与の視覚音素シーケンスに従って顔の口をアニメーション化するコンピュータ実施方法。

【請求項 2】

前記行列に基づいて主成分解析を実行して、前記複数の視覚音素の各々の前記主成分を取得する、実行するステップと、

前記サンプルに基づいて、前記フェイスアニメーションモデル内の前記複数の V 個の視覚音素の各々の前記統計学的変動性を表す統計学的モデルを取得することステップと、

前記視覚音素空間を通る前記軌跡の平滑性の確率的表現により、或る視覚音素表現の前記事前確率を表現するステップと、

前記事前確率及び前記統計学的モデルに基づいて、前記ベイズ手法を適用して、前記所与の視覚音素シーケンスの前記視覚音素空間を通る最良のパスを取得する、適用するステップと、

を更に含む、請求項 1 に記載のコンピュータ実施方法。

【請求項 3】

前記複数の I 個のサンプルは、異なる被写体により話された或る視覚音素のサンプルであるか、又は

前記複数の I 個のサンプルは、異なる言葉に属するか又は一人の被写体が異なる時点において話した或る視覚音素のサンプルである、請求項 1 ～ 3 のいずれか一項に記載のコンピュータ実施方法。

【請求項 4】

視覚音素 v について位置合わせされた人物 i のスキャンが x_{v_i} として示され、対応するニュートラルフリースキャン

【数 3】

$$\xi_i^v = x_i^v - x_i^{neut}$$

10

20

30

40

50

は、同じ個人 I のニュートラルスキャン

【数 4】

$$x_i^{\text{neutral}}$$

を差し引くことにより計算され、全ての視覚音素の前記ニュートラルフリースキャンが前記データ行列に積み重ねられる、請求項 1～4 のいずれか一項に記載のコンピュータ実施方法。

【請求項 5】

前記フェイスアニメーションモデル内のスキャンされた前記各視覚音素の前記変動性を記述する統計学的モデルを取得するステップは、 10

【数 5】

$$\beta_{i,v} = \text{diag}(\sigma_{iv}) V_T \xi_{vi}$$

により特定の視覚音素 v の全てのスキャン i を前記フェイスアニメーションモデルに射影するステップと、

好ましくは再び PCA を使用して、前記データポイントに多変数正規分布をフィッティングするステップと、

を含む、請求項 1～4 のいずれか一項に記載のコンピュータ実施方法。 20

【請求項 6】

前記各視覚音素の前記統計学的モデル

【数 6】

$$\mathcal{V} = (\text{diag}(\sigma_i), B, \bar{\beta} \in R^n)$$

は、前記 F A M

【数 7】

$$(\bar{\beta} \in R^n)$$

内の平均、成分の（列）正規直交行列 B、及び加重 $\text{diag}(\sigma_i)$ により定義される、請求項 1～5 のいずれか一項に記載のコンピュータ実施方法。

【請求項 7】

時点 t_i における視覚音素

【数 8】

$$v_i$$

のリスト

【数 9】

$$\mathcal{L} = \{v_i, t_i\}$$

をアニメーション化する場合、前記アニメーションは、フェイスアニメーションモデル

【数 10】

$$\beta(t) : R \rightarrow R^n$$

を通じて平滑な軌跡をフィッティングすることにより生成され、前記フェイスアニメーシ 50

ョンモデルは、頂点空間内で形状シーケンス

【数 1 1】

$$\xi(\beta(t)) = \text{diag}(\sigma_i^v) V \beta(t)$$

を定義する、請求項 1 ～ 6 のいずれか一項に記載のコンピュータ実施方法。

【請求項 8】

【数 1 2】

$$\prod p(\mathcal{V}_i | \beta(t_i)) = \prod \dots e^{-\frac{1}{2} \|\text{diag}(\sigma_i)^j B^j T (\beta(t_i) - \bar{\beta}^i)\|^2} \quad 10$$

である、請求項 1 ～ 7 のいずれか一項に記載のコンピュータ実施方法。

【請求項 9】

【数 1 3】

$$\begin{aligned} p(\beta) &= \prod p(\beta(t_i)) \\ &= \prod \eta_1 \dots e^{-\frac{1}{2} \|\nabla \xi(\beta(t_i))\|^2} + \eta_2 \dots e^{-\frac{1}{2} \|\nabla^2 \xi(\beta(t_i))\|^2} \end{aligned} \quad 20$$

である、請求項 1 ～ 8 のいずれか一項に記載のコンピュータ実施方法。

【請求項 1 0】

前記確率を最大化するために、前記費用関数

【数 1 4】

$$CF = \log p(\mathcal{V} | \beta) \cdot p(\beta) = \log p(\mathcal{V} | \beta) + \log p(\beta)$$

は最小化される、請求項 5 ～ 9 のいずれか一項に記載の方法。

30

【請求項 1 1】

所与の視覚音素シーケンスに従って顔の口をアニメーション化する装置であって、

フェイスモデルにより前記顔をグラフィックで表現するモジュールと、

複数の V 個の可能な視覚音素の各々に、該視覚音素の複数の I 個の異なるスキャン又はサンプルを測定により取得するモジュールと、

前記フェイスアニメーションモデルにより前記複数の V 個の異なる視覚音素の各々の前記複数の I 個のサンプルの各々を表現して、視覚音素シーケンスを視覚音素空間を通る軌跡を通じて表現することができるように、前記視覚音素空間に広がる前記スキャン又はサンプルに基づいて行列を生成するモジュールと、

ベイズ手法を適用して、前記所与の視覚音素シーケンスの前記視覚音素空間を通る最良のパスを取得するモジュールと、

40

を備える、装置。

【請求項 1 2】

請求項 2 ～ 1 0 のいずれか一項に記載の方法を実行するモジュールを更に備える、請求項 1 1 に記載の装置。

【請求項 1 3】

コンピュータで実行されると、前記コンピュータが請求項 1 ～ 1 0 のいずれか一項に記載の方法を実行できるようにするコンピュータ実行可能なコードを含むコンピュータプログラム。

【発明の詳細な説明】

50

【技術分野】

【0001】

本発明は、顔の口をアニメーション化するコンピュータ実施方法及び対応する装置に関する。

【0002】

一実施の形態では、本発明は、人のような仮想アバターの分野に関する。より詳細には、本発明は、発話アバター（speaking avatar）の口アニメーションの現実性及び自然性を増大させる。そのような人のようなアバターは、オンライン仮想アシスタント、仮想教師、パーソナルアバター（例えば、ビデオチャットで人物に取って代わる）、そして更に多くのものとすることができる。

10

【背景技術】

【0003】

キャラクタのアニメーション化は、コンピュータグラフィックス領域での困難な技術分野である。キャラクタのアニメーション化は、アニメーション化された動画を生成する際に使用されるが、動画のみならず、いわゆるアバターをアニメーション化する際にも使用することができる。

【0004】

仮想アバターは、オンラインアシスタント又はe-教師等の多くの用途で広く使用されている。仮想アバターは、発話及び体の動きを通じてそれ自体を表現して、情報を現実的な方法で伝える。アニメーションの表現性及び自然性により、これらのアバターはユーザから高く評価されるようになる。したがって、ユーザは、サービスが自然で良質なアバターを提供する場合、そのサービスの使用を高く評価する。

20

【0005】

自然な発話表現を提供するためには、唇（すなわち、発話の視覚的部分）は発話（すなわち、音声）と同期しなければならない、自然にアニメーション化されなければならない。このプロセスは通常、2つのステップで行われる：

- ・視覚音素（viseme）に対応するキーフレーム（すなわち、意味を区別する音の最小単位である視覚的な音素（visual phoneme））を定義する、
- ・キーフレーム間のアニメーションを補間又は実行する。

【0006】

30

キーフレームの定義及び音声との同期は、音声と視覚とのいかなる不連続性も回避するために重要である。第2のステップは、信頼性が高く自然なアニメーションを提供し、いかなる視覚的な不連続性も回避するためにも重要である。これらの不連続性（すなわち、音声と視覚との不連続性及び視覚的な不連続性）は、アバターの現実性レベル及び自然性を低下させ、ユーザを困惑させる。これは、自然ではないアバターを使用する場合、ユーザによるサービスの使用回避に繋がりが得る。

【0007】

顔のアニメーションは数年来、活発な研究分野であり、大規模な調査については、Deng及びNoh「Computer Facial Animation: A Survey」、（Deng及びNeumann「Data-Driven 3D Facial Animation」掲載）を参照のこと。既存の解決策は、ヒューリスティック手法、サンプルベースの手法、及び学習ベースの手法に分類することができる。

40

【0008】

ヒューリスティック手法は、手動で作成された視覚音素形状及び経験的な平滑化関数を使用する。サンプルベースの手法は、「トライフォンビデオセグメント」を取得して使用し、同時調音をモデリングする。

【0009】

学習ベースの手法は、データから視覚音素及び同時調音を学習する。例えば、BrandはHMMベースモデルを使用する（Brand、「Voice pupperty」、SIGGRAPH 99、参照）。Ezzat、Geiger及びPoggio「Trainable videorealistic speech animation」、ACM Trans.Graph., 388-398, 2002、では、多次元モーフィングモデルを使用する。視覚音素ごとに1ス

50

キャンを使用する視覚音素モデルを使用することが、例えばV.Blanz、C.Basso、T.Poggio及びT.Vetter、「Reanimating Faces in Images and Video」、EUROGRAPHICS 2003、に開示されている。Deng、Neumann、Lewis、Kim、Bulut、Narayanan、「Expressive facial animation synthesis by learning speech co-articulations and expression spaces」、IEEE Transaction on Visualization and Computer Graphics,2006、では、同時調音モデルの学習が開示されている。

【発明の概要】

【発明が解決しようとする課題】

【0010】

本発明の目的は、発話アバターの口のアニメーションの現実性及び自然性を増大させるように口のアニメーションを改良することである。そのような人のようなアバターは、オンライン仮想アシスタント、仮想教師、パーソナルアバター（例えば、ビデオチャットで人物に取って代わる）、そして更に多くのものとすることができる。

【課題を解決するための手段】

【0011】

本発明の一実施の形態によれば、所与の視覚音素シーケンスに従って顔の口をアニメーション化するコンピュータ実施方法であって、

フェイスモデルにより前記顔をグラフィックで表現するステップと、

複数のV個の可能な各視覚音素に、該視覚音素の複数のI個の異なるスキャン又はサンプルを測定により取得するステップと、

前記フェイスアニメーションモデルにより前記複数のV個の異なる各視覚音素の前記複数のI個の各サンプルを表現して、視覚音素シーケンスを視覚音素空間を通る軌跡を通じて表現することができるように、前記視覚音素空間に広がる前記スキャン又はサンプルに基づいて行列を生成する、表現するステップと、

ベイズ手法を適用して、前記所与の視覚音素シーケンスの前記視覚音素空間を通る最良のパスを取得する、適用するステップと、を含む、所与の視覚音素シーケンスに従って顔の口をアニメーション化するコンピュータ実施方法が提供される。

【0012】

このようにして、視覚音素表現の生じ得る変動性を考慮に入れることができ、それにより、より自然なアニメーションが達成される。

【0013】

一実施の形態によれば、前記方法は、前記行列に基づいて主成分解析を実行して、前記複数の各視覚音素の前記主成分を取得する、実行するステップと、

前記サンプルに基づいて、前記フェイスアニメーションモデル内の前記複数のV個の各視覚音素の前記統計学的変動性を表す統計学的モデルを生成するステップと、

前記視覚音素空間を通る前記軌跡の平滑性の確率的表現により、或る視覚音素表現の前記事前確率を表現するステップと、

前記事前確率及び前記統計学的モデルに基づいて、前記ベイズ手法を適用して、前記所与の視覚音素シーケンスの前記視覚音素空間を通る最良のパスを取得する、適用するステップと、を更に含む。

【0014】

統計学的モデル及び平滑性基準により、ベイズモデルの適用及び実施が可能である。

【0015】

一実施の形態によれば、前記複数のI個のサンプルは、異なる被写体により話された或る視覚音素のサンプルであるか、又は

前記複数のI個のサンプルは、異なる言葉に属するか又は一人の対象が異なる時点において話した或る視覚音素のサンプルである。

【0016】

10

20

30

40

50

これにより、視覚音素の変動性をモデルで考慮することが可能である。

【0017】

一実施の形態によれば、視覚音素 v について位置合わせされた人物 i のスキャンが x_v i として示され、対応するニュートラルフリースキャン

【数1】

$$\xi_i^v = x_i^v - x_i^{\text{neutral}}$$

は、同じ個人 I のニュートラルスキャン

【数2】

$$x_i^{\text{neutral}}$$

10

を差し引くことにより計算され、全ての視覚音素の前記ニュートラルフリースキャンが前記データ行列に積み重ねられる。

【0018】

これは、視覚音素形状を抽出し、形状へのニュートラルスキャンの影響をなくす。

【0019】

一実施の形態によれば、前記フェイスアニメーションモデル内のスキャンされた前記各視覚音素の前記変動性を記述する統計学的モデルを取得するステップは、

20

【数3】

$$\beta_{i,v} = \text{diag}(\sigma_i^v) V^T \xi_i^v$$

により特定の視覚音素 v の全てのスキャン i を前記フェイスアニメーションモデルに射影するステップと、

好ましくは再びPCAを使用して、前記データポイントに多変数正規分布をフィッティングするステップと、

を含む。

【0020】

30

これにより、統計学的モデルの実施が可能になる。

【0021】

一実施の形態によれば、前記各視覚音素の前記統計学的モデル

【数4】

$$\mathcal{V} = (\text{diag}(\sigma_i), B, \beta \in R^n)$$

は、前記FAM

【数5】

$$(\bar{\beta} \in R^n)$$

40

内の平均、成分の(列)正規直交行列 B 、及び加重 $\text{diag}(\sigma_i)$ により定義される。

【0022】

一実施の形態によれば、時点 t_i における視覚音素

【数6】

$$V_i$$

のリスト

50

【数 7】

$$\mathcal{L} = (\mathcal{V}_i, t_i)$$

をアニメーション化する場合、前記アニメーションは、フェイスアニメーションモデル

【数 8】

$$\beta(t) : R \rightarrow R^n$$

を通じて平滑な軌跡をフィッティングすることにより生成され、前記フェイスアニメーションモデルは、頂点空間内で形状シーケンス 10

【数 9】

$$\xi(\beta(t)) = \text{diag}(\sigma_i^u) \mathbf{V} \beta(t)$$

を定義する。

【0 0 2 3】

これにより、最適な軌跡の導出が可能になる。

【0 0 2 4】

一実施の形態によれば、前記視覚音素リストを所与とした前記軌跡の確率 20

【数 1 0】

$$p(\beta|\mathcal{L}) \sim p(\mathcal{L}|\beta) \cdot p(\beta)$$

は、前記軌跡を所与とした前記視覚音素リストの確率

【数 1 1】

$$p(\mathcal{L}|\beta) = \prod p(\mathcal{V}_i|\beta(t_i))$$

に比例する。 30

【0 0 2 5】

これはベイズ理論の実施である。

【0 0 2 6】

一実施の形態によれば、前記軌跡を所与とした前記視覚音素リストの確率は、

【数 1 2】

$$\prod p(\mathcal{V}_i|\beta(t_i)) = \prod \dots e^{-\frac{1}{2} \|\text{diag}(\sigma_i)^u \mathbf{B}^u \mathbf{J}^T (\beta(t_i) - \beta^u)\|^2}$$

【0 0 2 7】

これにより、確率分布の実施が可能になる。 40

【0 0 2 8】

一実施の形態によれば、

【数 1 3】

$$\begin{aligned} p(\beta) &= \prod p(\beta(t_i)) \\ &= \prod \eta_1 \dots e^{-\frac{1}{2} \|\nabla \xi(\beta(t_i))\|^2} + \eta_2 \dots e^{-\frac{1}{2} \|\nabla^2 \xi(\beta(t_i))\|^2} \end{aligned}$$

である。

【0 0 2 9】

一実施の形態によれば、前記確率を最大化するために、前記費用関数 50

【数 1 4】

$$CF = \log p(V|\beta) \cdot p(\beta) = \log p(V|\beta) + \log p(\beta)$$

は最小化される。

【0030】

一実施の形態によれば、所与の視覚音素シーケンスに従って顔の口をアニメーション化する装置であって、

フェイスモデルにより前記顔をグラフィックで表現するモジュールと、

複数の V 個の可能な各視覚音素に、該視覚音素の複数の I 個の異なるスキャン又はサンプルを測定により取得するモジュールと、

前記フェイスアニメーションモデルにより前記複数の V 個の異なる各視覚音素の前記複数の I 個の各サンプルを表現して、視覚音素シーケンスを視覚音素空間を通る軌跡を通じて表現することができるように、前記視覚音素空間に広がる前記スキャン又はサンプルに基づいて行列を生成するモジュールと、

ベイズ手法を適用して、前記所与の視覚音素シーケンスの前記視覚音素空間を通る最良のパスを取得するモジュールと、

を備える、所与の視覚音素シーケンスに従って顔の口をアニメーション化する装置が提供される。

【0031】

このようにして、本発明の一実施形態による装置を実施することができる。

【0032】

一実施形態によれば、方法が提供され、本発明の実施形態の 1 つによる方法を実行するモジュールを備える装置が提供される。

【0033】

一実施の形態によれば、コンピュータで実行されると、前記コンピュータが本発明の実施の形態のうちの 1 つによる方法を実行できるようにするコンピュータ実行可能なコードを含むコンピュータプログラムが提供される。

【図面の簡単な説明】

【0034】

【図 1】本発明の実施形態を概略的に示す図である。

【図 2】本発明の実施形態を概略的に示す図である。

【図 3】本発明の実施形態を概略的に示す図である。

【図 4】本発明の実施形態を概略的に示す図である。

【発明を実施するための形態】

【0035】

以下、本発明の幾つかの実施形態について説明する。

【0036】

一実施形態によれば、発話に従ってアバターの口をアニメーション化する方法は、

・視覚音素ごとに 1 組の可能な形状を取得し、

・動きの成分を通じて、視覚音素空間内の全ての視覚音素について全ての可能な口の動きを生成することができる生成統計学的モデルであるフェイスアニメーションモデル (FAM) を作成する、

・FAM 視覚音素空間内の視覚音素間で統計的に最良のパスを見つける、
である。

【0037】

視覚音素ごとに複数の形状を作成することにより、各視覚音素の表現に柔軟性が提供される。実際には、視覚音素の表現は、異なる事例で (例えば、人ごと、かつ、言葉ごとに) 異なり得る。したがって、視覚音素ごとのこのより広い 1 組の形状によって、より大きな 1 組の可能性を有することができる。

【0038】

視覚音素は異なる方法で表現することができるため（すなわち、異なる値を仮定する視覚音素空間内の成分による僅かな違いを有する）、最も適切に表現された視覚音素（成分の「最良値」）を識別する選択を行う必要がある。それを達成するために、生成統計学的モデルが作成される。このモデル（すなわち、FAM）は、視覚音素の全ての可能な形状間の全ての可能な口の動きを生成可能である。これは、フェイスアニメーションモデルを作成するステップにおいて作成された形状セットに基づいて、全ての可能な口の動きを含む空間として視覚化することができる。FAMの作成後、次のステップは、シーケンスの個々の視覚音素を表す形状間を自然に補間するために、この空間内の最良パスを見つけることである。

10

【0039】

視覚音素ごとに1つの形状を使用するこれまでの手法とは対照的に、この手法では、各視覚音素の口形状の統計学的3Dモデルが使用される。このモデルは、異なる被写体（又は別の実施形態では、同じ視覚音素を含む異なる言葉）にわたる特定の視覚音素の口形状の変動性を符号化する。この手法の主要な利点は、はるかに自然な口のアニメーション化が可能なことである。一実施形態によれば、これにより、時点tにおいて抽出された視覚音素に属する口形状の尤度と、軌跡の平滑性についての先験的仮定とをトレードオフして、ベイズ手法を使用してキーフレーム間を補間することができる。

【0040】

以下、更なる実施形態について幾らかより詳細に説明する。

20

【0041】

以下に説明される実施形態は、モーフィングモデル（MM）と呼ばれ、V.Blanz及びT.Vetter、「A Morphable Model for the Synthesis of 3D Faces」、SIGGRAPH 99,187-194、に記載されている特定の3Dフェイスモデルを利用する。しかし、視覚音素がその視覚音素を表す個々の成分からなる視覚音素空間内にアニメーション化された視覚音素を表すことが可能な任意のモデルを使用することができる。

【0042】

この実施形態では、特定の人物の口（歯を含む）の現実的でもっともらしいアニメーションを生成可能なフェイスアニメーションシステムについて説明する。口の動きのために、データから学習される口形状の統計学的モデル（すなわち、13個の各視覚音素に1つの統計学的モデル）が使用される。

30

【0043】

このために、幾つかのステップが実行される。

【0044】

第1のステップにおいて、視覚音素が獲得又は取得される。

【0045】

視覚音素は、既知のスキャン装置を用いて取得される。一実施形態によれば、アイデンティティ（identity）ごと（すなわち、人物ごと）に、16個のスキャン（13個の視覚音素及び基準として口が閉じられた状態での3つのニュートラルスキャン）が取得される。実施形態では、12のアイデンティティの視覚音素がスキャンされ、合計で192のスキャンが得られる。スキャン又は人物の数に他の値を選ぶことも同様に可能である。

40

【0046】

次のステップにおいて、フェイスアニメーションモデルが生成される。フェイスアニメーションモデル（FAM）は、全ての視覚音素の全ての可能な口の動きを生成可能な生成統計学的モデルである。FAMはモーフィングモデル（MM）と一緒に使用される：MMは、アイデンティティを符号化し、3D再構築の計算に使用される。FAMは、口の動きを符号化し（アイデンティティに影響を及ぼさない）、フェイスアニメーションに使用される。

【0047】

FAMを完了するために、まず、スキャンを対応させ、これは、この例で図1の3行目

50

に示されるように、全てのスキャンの共通パラメータ化（すなわち、共通メッシュ）を定義する。図 1 は、3つのアイデンティティ（1列目、2列目、3列目のそれぞれ）の視覚音素 a o の幾つかのスキャン例を示す。2列目は未処理スキャンを示し、3列目は処理済みデータを示す。

【0048】

視覚音素 v の人物 i の位置合わせされたスキャンが x_{v_i} として示され、対応するニュートラルフリースキャン

【数 15】

$$\xi_i^v = x_i^v - x_i^{\text{neutral}}$$

10

は、同じ個人 i のニュートラルスキャン

【数 16】

$$x_i^{\text{neutral}}$$

を差し引くことにより計算される。全ての視覚音素のニュートラルフリースキャンがデータ行列

【数 17】

$$X = [\xi_1^{ao}, \xi_2^{ao}, \dots, \xi_1^v, \dots, \xi_i^v, \dots]$$

20

に積み重ねられる。

【0049】

モデルの成分を計算するために、主成分解析（PCA）が1組のスキャン例に対して使用される。

【数 18】

$$X = VWY$$

30

式中、成分 $V = [v_1, \dots, v_n]$ は（列）正規直交成分であり、 W は対角成分であり、 $\sigma_{i_v} = w_i / \sqrt{n}$ であり、式中、 n はスキャン数である。結果として生成される、データから学習されるモデルは図 2 において視覚化される。

【0050】

上述したように、フェイスアニメーションモデル F A M は MM と一緒に使用される：各アニメーション化顔形状

【数 19】

$$x = \underbrace{\mu + \sum \alpha_i \sigma_i^u u_i}_{\text{Fitting}} + \underbrace{\sum \beta_i \sigma_i^v v_i}_{\text{Fitting}} = \underbrace{\mu + \xi_{id}}_{\text{Fitting}} + \xi_u$$

40

は、MM 平均 μ 、単位成分 u_i の線形結合（MM 及びフィッティングにより与えられる）、及びアニメーションの成分の線形結合（加重 σ_{i_v} とともに前に計算された v_i ）として表現される。次に、各形状が、単位系数 α_i 及びそのアニメーション系数 β_i により符号化される。

【0051】

次のステップにおいて、発話アニメーションの場合、フェイスアニメーションモデル内でスキャンされた13個の各視覚音素の変動性を記述する統計学的モデルが取得される。

そのようなモデルを計算するために、特定の視覚音素 v の全てのスキャン i が F A M に射

50

影され、

【数 2 0】

$$\beta_{i,v} = \text{diag}(\sigma_i^v) V^T \xi_i^v$$

再び P C A を使用して、データポイントに多変数正規分布がフィッティングされる。各視覚音素の統計学的モデル

【数 2 1】

$$\mathcal{V} = (\text{diag}(\sigma_i), B, \beta \in R^n)$$

10

は、F A M

【数 2 2】

$$(\bar{\beta} \in R^n)$$

内の平均、成分の（列）正規直交行列 B、及び加重 $\text{diag}(\sigma_i)$ により定義される。そのような視覚音素モデルの例は図 3 に見ることができ、図 3 は、視覚音素 a o（上）及び f（下）の視覚音素モデル例を示す。

【0 0 5 2】

20

このようにして、視覚音素の視覚音素空間の成分の「確率分布」を取得することができる。そうして得られた統計学的モデルは、上記サンプルに基づいて上記フェイスアニメーションモデル内で複数の V 個の視覚音素のそれぞれの統計学的変動性を表す。

【0 0 5 3】

以下、一実施形態によるアニメーションについて説明する。

【0 0 5 4】

この実施形態では、視覚音素取得時に得られたタイミング情報を視覚音素モデルと共に使用して、事前に記録された音ファイルに従って画像アニメーションをレンダリングする。このタイミング情報は、書かれた言葉に基づいて、又は何らかの方法で得られた連結された視覚音素のみに基づいて導出された視覚音素シーケンスのレンダリングにも同様に使用することができる。

30

【0 0 5 5】

発話解析の出力は、時点 t_i における視覚音素

【数 2 3】

$$\mathcal{V}_i$$

のリスト

【数 2 4】

$$\mathcal{L} = (\mathcal{V}_i, t_i)$$

40

である。アニメーションは、F A M

【数 2 5】

$$\beta(t) : R \rightarrow R^n$$

を通じて平滑な軌跡をフィッティングすることにより生成され、この F A M は、頂点空間内で形状シーケンス

【数 2 6】

$$\xi(\beta(t)) = \text{diag}(\sigma_t^v) V \beta(t)$$

を定義する。

【0056】

フィッティングはベイズ手法を使用して行われる。これは図4に示され、図4は、発話アニメーションの場合、フェイスアニメーションモデルを通じて平滑な軌跡をどのようにしてフィッティングするかを示す。時点 t において、対応する視覚音素のもっともらしい形状が生成される（3つの異なる視覚音素の統計学的モデルが可視化される）（図4参照）。

【0057】

ベイズ手法は、その時点における視覚音素に属するキーフレーム f 内の形状の確率と、一次導関数及び二次導関数を使用する口の動きの平滑性の先験的仮定とのトレードオフを含む。

【0058】

換言すれば、平滑性の確率的表現により表される軌跡の平滑性についての先験的仮定が利用される。これは、特定の口の動き（或る形状から他の形状への遷移を表す）の速度及び加速度が、この速度及び加速度が幾らか制限されるように或る確率分布に従うという仮定に基づく。

【0059】

視覚音素リストを所与とした軌跡の確率

【数 2 7】

$$p(\beta | \mathcal{L}) \sim p(\mathcal{L} | \beta) \cdot p(\beta)$$

は、軌跡を所与とした視覚音素リストの確率

【数 2 8】

$$p(\mathcal{L} | \beta) = \prod p(\mathcal{V}_t | \beta(t_t)) \quad (2.1) \quad 30$$

$$= \prod \dots e^{-\frac{1}{2} \|\text{diag}(\sigma_t) V^T (\beta(t_t) - \beta^*)\|^2} \quad (2.2)$$

に比例し（これはベイズ理論の適用）、ここでは、前に得られた統計学的視覚音素モデル、及び軌跡の先験的確率

【数 2 9】

$$p(\beta) = \prod p(\beta(t_i)) \quad (2.3)$$

$$= \prod \eta_1 \dots e^{-\frac{1}{2} \|\nabla \xi(\beta(t_i))\|^2} + \eta_2 \dots e^{-\frac{1}{2} \|\nabla^2 \xi(\beta(t_i))\|^2} \quad (2.4) \quad 40$$

$$= \prod \eta_1 \dots e^{-\frac{1}{2} \|\xi(\beta(t_i)) - \xi(\beta(t_{i+1}))\|^2} \quad (2.5)$$

$$+ \eta_2 \dots e^{-\frac{1}{2} \|\xi(\beta(t_{i+1})) - 2\xi(\beta(t_i)) + \xi(\beta(t_{i-1}))\|^2} \quad (2.6)$$

が利用され、ここでは、軌跡の平滑性についての先験的仮定、特に、人が口を素早く動かすことができず（一次導関数＝速さ、パラメータ η_1 ）、素早く加速させることができない（二次導関数＝速度：パラメータ η_2 ）という先験的仮定がなされる。

【0060】

パラメータ η_1 、パラメータ η_2 は実験的に決定することができる。

【0061】

確率を最大にするために、L-BFGS-Bを使用して、標準ニュートンオプティマイザを使用して費用関数

【数 3 0】

$$CF = \log p(\mathcal{V}|\beta) \cdot p(\beta) = \log p(\mathcal{V}|\beta) + \log p(\beta) \quad (2.7)$$

$$= \sum_{i=1}^M \|\text{diag}(\sigma_{i,j}) \mathbf{U}_i^T (\mathbf{s}(t_i) - \tilde{\mathbf{c}}_i)\|^2 + \quad (2.8)$$

$$\eta_1 \sum_{i=1}^{M-1} \|\mathbf{x}(\beta(t_i)) - \mathbf{x}(\beta(t_{i+1}))\|^2 + \quad (2.9)$$

$$\eta_2 \sum_{i=2}^{M-1} \|\mathbf{x}(\beta(t_{i+1})) - 2\mathbf{x}(\beta(t_i)) + \mathbf{x}(\beta(t_{i-1}))\|^2 \quad (2.10)$$

10

が最小化され、Mはフレーム数である。上記式 2. 8 中、以下の修正を行うべきである：UはWを示し、s (t i) は

【数 3 1】

$$\mathbf{x}(\beta(t_i))$$

を示し、

【数 3 2】

$$\tilde{\mathbf{c}}_i'$$

20

は

【数 3 3】

$$\beta$$

を示すべきである。

【0 0 6 2】

上記例の代わりに、同様に他の最適化関数を使用することもできる。

【0 0 6 3】

30

このようにして、視覚音素空間を通る最適な軌跡を与える視覚音素表現を取得することができ、口アニメーションのレンダリングに使用することができる。

【0 0 6 4】

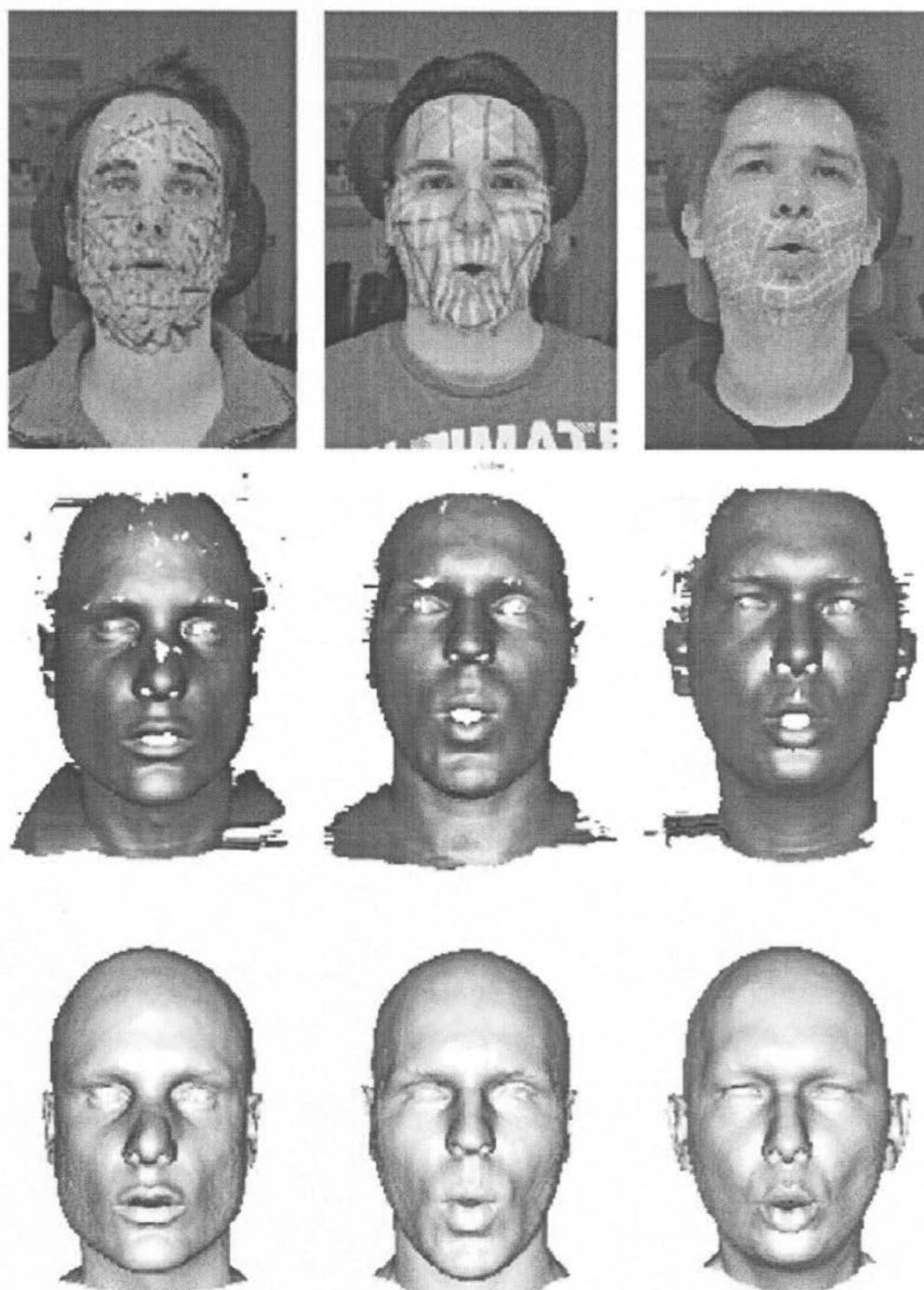
最後のビデオは、結果として生成される形状をレンダリングすることにより、当該技術分野で既知のワープフィールド／3D組み合わせ手法を使用してレンダリングされる。

【0 0 6 5】

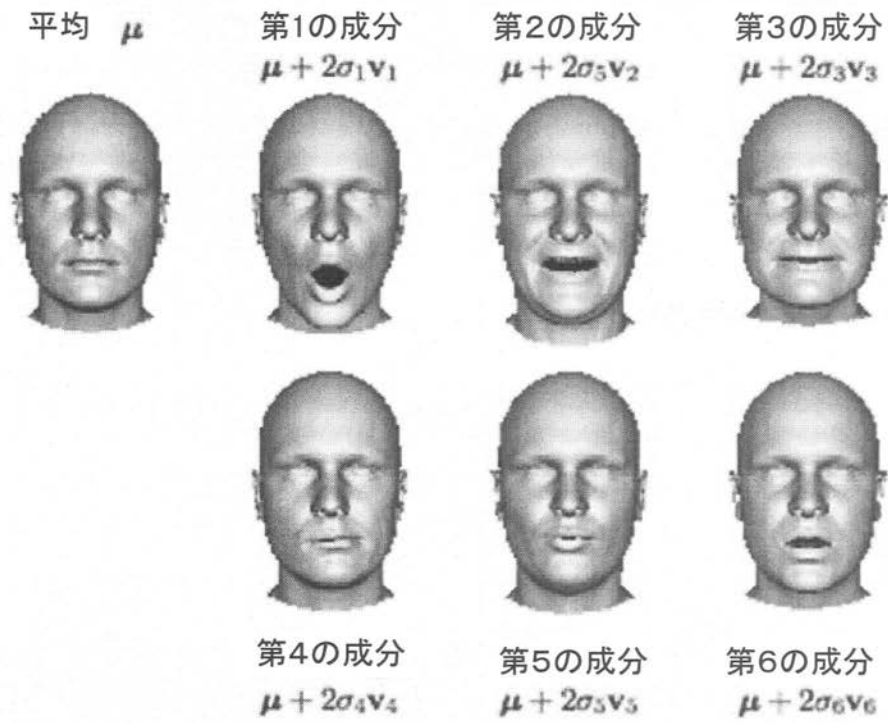
本発明の実施形態に関連して説明された方法、要素、ユニット、及び装置をハードウェア、ソフトウェア、又は両方の組み合わせとして実施可能なことが当業者には容易に明らかであろう。特に、本発明の実施形態及び実施形態に関連して説明されたモジュールの要素が、コンピュータで実行されるか、又はマイクロプロセッサにより実行される単数又は複数のコンピュータプログラムにより実施可能なことが理解されよう。本発明を実施する任意の装置は特に、ネットワーク内で動作するルータ、サーバ、モジュール等のネットワークエンティティの形態、又は携帯電話、スマートフォン、PDA等のモバイル装置の形態、又は同様の任意のものの形態をとることができる。

40

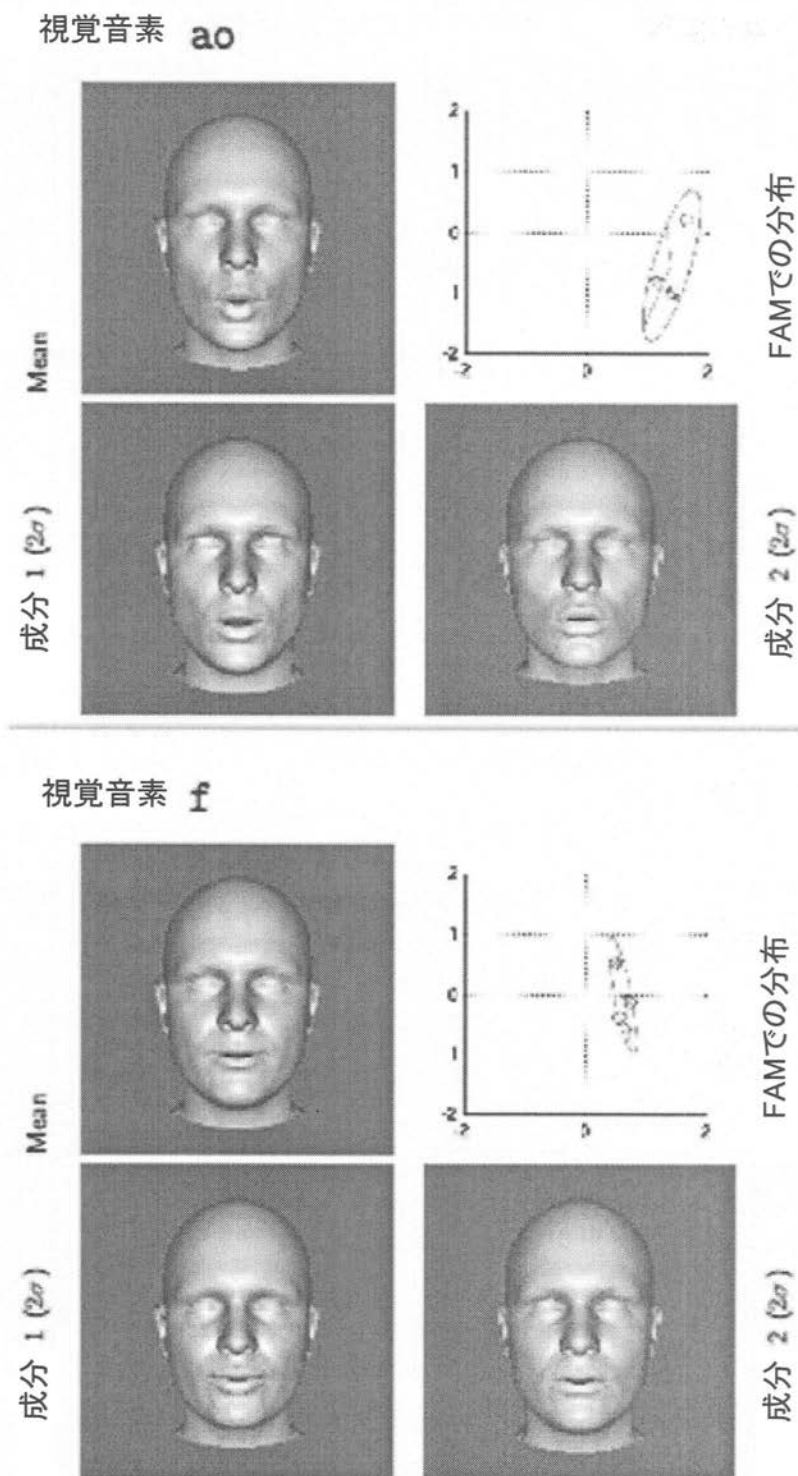
【図 1】

**Fig. 1**

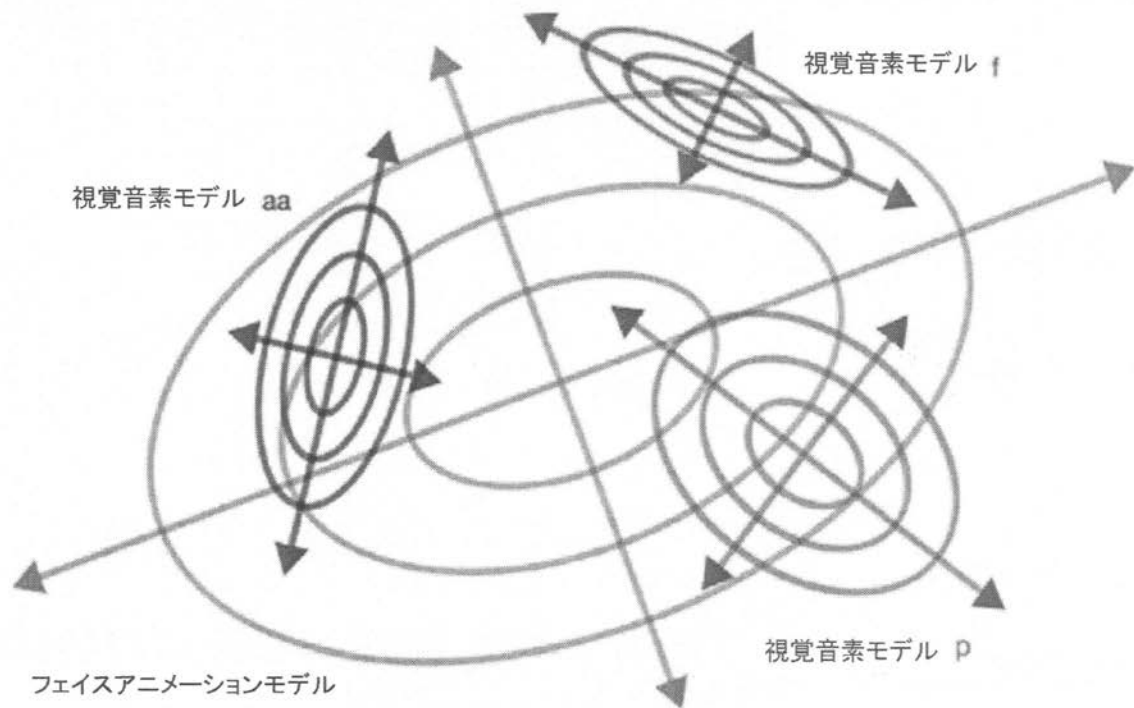
【図 2】

**Fig. 2**

【図 3】

**Fig. 3**

【図 4】

**Fig. 4**

フロントページの続き

(74)代理人 100154298

弁理士 角田 恭子

(74)代理人 100166268

弁理士 田中 祐

(74)代理人 100170379

弁理士 徳本 浩一

(74)代理人 100161001

弁理士 渡辺 篤司

(72)発明者 ラインハルト・クノーテ

スイス国, 4 0 5 8 バーゼル, シェーナウシュトラーセ 8 4

(72)発明者 トーマス・フェッター

スイス国, 4 0 5 7 バーゼル, アマーバッハシュトラーセ 7 8

(72)発明者 ラミ・アジャジ

ドイツ連邦共和国, 8 0 7 9 8 ミュンヘン, ルイーゼンシュトラーセ 6 0 アー

(72)発明者 ミヒャエル・ファールマイアー

ドイツ連邦共和国, 8 1 3 7 5 ミュンヘン, グアルディーニシュトラーセ 1 7 0

Fターム(参考) 5B050 BA08 BA09 BA12 DA01 EA07 EA13 EA19 EA24 EA27 FA02

FA10

【外国語明細書】

**COMPUTER-IMPLEMENTED METHOD AND APPARATUS
FOR ANIMATING A MOUTH OF A FACE****5 FIELD OF INVENTION**

The present invention relates to a computer-implemented method for animating a mouth of a face and a corresponding apparatus.

- 10 In one embodiment it relates to the field of human-like virtual avatars. More specifically, it increases the realism and the naturalness of the mouth animation of the speaking avatar. Such human-like avatars could be online virtual assistants, virtual teachers, personal avatars (e.g. that replace persons in video chats) and many more.

15 BACKGROUND OF THE INVENTION

Animating characters is a challenging technological field in the area of computer graphics. It is used when producing animated movies, but also it can be used not only for movies but for animating so-called avatars.

- 20 Virtual avatars are widely used in many applications such as online assistants or as e-teachers. They express their selves through speech and body movement to relate the information in a realistic way. The expressivity and the naturalness of animations make these avatars appreciated by users. Therefore users appreciate using a service when it offers a natural and a
25 good quality avatar.

In order to provide a natural speech expression, the lips (i.e. the visual part of the speech) must be synchronized with the speech (i.e. the audio) and must be animated in a natural way. This process is usually done in two steps:

- 30
- Define keyframes corresponding to visemes (i.e. visual phoneme which is the smallest unit of sound that distinguishes meaning)
 - Interpolate or perform the animation between keyframes

- The definition of the keyframes as well as their synchronization with audio is important to avoid any audio-visual discontinuity. The second step is also important to offer a realistic and natural
35 animation and avoid any visual discontinuities. These discontinuities (i.e. the audio-visual and the visual ones) degrade the realism level and the naturalness of the avatar and will disturb the users. This may lead users to avoid using a service if it uses a non-natural avatar.

Facial Animation is an active research area since years, for an extensive survey one refer to Deng and Noh: Computer Facial Animation: A Survey (in: Deng and Neumann Data-Driven 3D Facial Animation). The existing solutions can be categorized into heuristic approaches, sample-based approaches and learning-based approaches.

5 Heuristic approaches use manually created viseme shapes and empirical smooth functions. Sample-base approaches acquire 'triphone video segments' and use them to model the coarticulation.

10 Learning-based approaches learn the visemes and coarticulation from data. For example, Brand uses a HMM-based model (see Brand: Voice puppetry. SIGGRAPH 99). In Ezzat, Geiger and Poggio: Trainable videorealistic speech animation. ACM Trans. Graph., 388-398, 2002, there is used a mutidimensional Morphable Model. To use a Viseme model with one scan per viseme is e.g. disclosed in V. Blanz, C. Basso, T. Poggio and T. Vetter, "Reanimating Faces in Images and Video" EUROGRAPHICS 2003. In Deng, Neumann, Lewis, Kim, Bulut, Narayanan. Expressive
15 facial animation synthesis by learning speech co-articulations and expression spaces, IEEE Transaction on Visualization and Computer Graphics, 2006, it is disclosed to learn co-articulation models.

20 It is an object of the present invention to improve mouth animation such that it increases the realism and the naturalness of the mouth animation of the speaking avatar. Such human-like avatars could be online virtual assistants, virtual teachers, personal avatars (e.g. that replace persons in video chats) and many more.

25 **SUMMARY OF THE INVENTION**

According to one embodiment there is provided a computer-implemented method of animating a mouth of a face in accordance with a given sequence of visemes, said method comprising:

Graphically representing said face by a face model;

30 For each of a plurality V of possible visemes, obtaining by measurement a plurality of I different scans or samples of said visemes;

Representing each of said plurality of I samples of each of said plurality of V different visemes by said face animation model to generate a matrix based on said scans or samples which spans up a viseme space such that a sequence of visemes can be represented through a trajectory

35 through said viseme space;

Applying a Bayesian approach to obtain the best path through said viseme space for said given sequence of visemes.

In this manner the possible variability of viseme representations can be taken into account to thereby achieve a more natural animation.

According to one embodiment the method further comprises:

- 5 Performing a principle component analysis based on said matrix to obtain the principle components for each of said plurality of visemes;
Generating a statistical model which represents the statistical variability of each of the plurality of V visemes within said face animation model based on said samples;
Representing the prior probability of a certain viseme representation by a probabilistic
- 10 representation of the smoothness of the trajectory through said viseme space;
Based on said prior probability and said statistical model, applying said Bayesian approach to obtain the best path through said viseme space for said given sequence of visemes.

- 15 The statistical model and the smoothness criterion allow for the application and implementation of a Bayes model.

- According to one embodiment Said plurality of I samples are samples of a certain visemes spoken by different subjects; or
said plurality of I samples are samples of a certain viseme belonging to different words or spoken
- 20 by a subject at different times.

This allows taking into account the variability of the visemes into the model.

- 25 According to one embodiment registered scan of person i for viseme v is denoted with x_{vi} , the corresponding neutral-free scan

$$\xi_i^v = x_i^v - x_i^{neutral}$$

- is computed by subtracting the neutral scan $x_i^{neutral}$ of the same individual. I, and
the neutral-free scans for all visemes are stacked to the data matrix

30

This extracts the viseme shape and eliminates the neutral scan effect on the shape.

- According to one embodiment said obtaining a statistical model that describes the variability of each of the visemes which have been scanned within the Face Animation Model comprises:
- 35 projecting all scans i of a specific viseme v into the Face Animation Model by

$$\beta_{i,v} = \text{diag}(\sigma_i^v) V^T \xi_i^v$$

and fitting a multivariate Gaussian distribution to the data points, preferably again using PCA.

- 5 This enables the implementation of a statistical model.

According to one embodiment the statistical models for each of the visemes

$$\mathcal{V} = (\text{diag}(\sigma_i), \mathbf{B}, \bar{\beta} \in R^n)$$

- 10 are defined by the mean within the FAM $(\bar{\beta} \in R^n)$, a (column-) orthonormal matrix of components B and a weighting $\text{diag}(\sigma_i)$.

According to one embodiment
for animating a list of viseme $\mathcal{V}_i$ at time t_i

15

$$\mathcal{L} = (\mathcal{V}_i, t_i)$$

the animation is generated by fitting a smooth trajectory through the
Face Animation Model

20

$$\beta(t) : R \rightarrow R^n$$

that defines a sequence of shapes

$$\xi(\beta(t)) = \text{diag}(\sigma_i') \mathbf{V} \beta(t)$$

25

in vertex space.

This enables the derivation of an optimum trajectory.

- 30 According to one embodiment the probability of the trajectory given the list of viseme

$$p(\beta|\mathcal{L}) \sim p(\mathcal{L}|\beta) \cdot p(\beta)$$

is proportional to the probability of the list of viseme given the trajectory

$$p(\mathcal{L}|\beta) = \prod p(\mathcal{V}_i|\beta(t_i))$$

5 This is an implementation of Bayes theorem.

According to one embodiment probability of the list of viseme given the trajectory

$$\prod p(\mathcal{V}_i|\beta(t_i)) = \prod \dots e^{-\frac{1}{2} \{\text{diag}(\sigma_i)^2 \mathbf{B}^T (\beta(t_i) - \beta^*)\}^2}$$

10

This allows the implementation of the probability distribution

According to one embodiment

$$\begin{aligned} p(\beta) &= \prod p(\beta(t_i)) \\ &= \prod \eta_1 \dots e^{-\frac{1}{2} \|\nabla \xi(\beta(t_i))\|^2} + \eta_2 \dots e^{-\frac{1}{2} \|\nabla^2 \xi(\beta(t_i))\|^2} \end{aligned}$$

15

According to one embodiment

to maximize the probability, there is minimized the cost function

$$CF = \log p(\mathcal{V}|\beta) \cdot p(\beta) = \log p(\mathcal{V}|\beta) + \log p(\beta)$$

20

According to one embodiment there is provided an apparatus for animating a mouth of a face in accordance with a given sequence of visemes, said apparatus comprising:

a module for graphically representing said face by a face model;

25 a module for, for each of a plurality V of possibly visemes, obtaining by measurement a plurality of I different scans or samples of said visemes;

a module for representing each of said plurality of I samples of each of said plurality of V different visemes by said face animation model to generate a matrix based on said scans or samples which spans up a viseme space such that a sequence of visemes can be represented through a

30 trajectory through said viseme space;

a module for applying a Bayesian approach to obtain the best path through said viseme space for said given sequence of visemes.

In this way an apparatus according to an embodiment of the invention may be implemented.

35

According to one embodiment there is provided an apparatus
a module for carrying out a method according to one of the embodiments of the invention.

- 5 According to one embodiment there is provided computer program comprising computer executable code which when being executed on a computer enables said computer to carry out a method to one of the embodiments of the invention.

DESCRIPTION OF THE DRAWINGS

10

Figures 1 to 4 schematically illustrated embodiments of the invention.

DETAILED DESCRIPTION

15

In the following several embodiments of the invention will be described.

According to one embodiment the method to animate the mouth of the avatar accordingly to the speech is:

20

- obtaining a set of possible shapes for each viseme
- Create a Face Animation Model (FAM) which is a generative statistical model that is able to generate all possible mouth movements for all visemes in a visems space through components of movement
- Find the statistically best path between visemes in the FAM viseme space

25

The creation of multiple shapes for each viseme offers flexibility for the expression of each viseme. Indeed, the expression of a viseme can vary in different cases (e.g. from a person to another and from a word to another). Therefore, this wider set of shapes for each viseme allows having a larger set of possibilities.

30

Since a viseme can be expressed in different ways (i.e. with slight variations by the components in the viseme space assuming different values), a choice has to be made for identifying the most suitable expressed viseme (the "best values" for the components). In order to accomplish that, a generative statistical model is created. This model (i.e. the FAM) is able to generate all possible mouth movements between all possible shapes of visemes. It can be visualized as a space enclosing all possible mouth movements based on the sets of shapes which have been created in the step of creating the face animation model. After the creation of the FAM, the next step is to find the best path in this space in order to naturally interpolate between the shapes representing

35

the individual visemes of a sequence.

In contrast to previous approaches that use a single shape per viseme, in this approach a statistical 3D model of the mouth shape for each viseme is used. This model codes the variability
5 of the mouth shape for a specific viseme across different subjects (or, in another embodiment, across different words containing the same viseme). The key advantage of this approach is that it allows a much more natural mouth animation: It allows, according to one embodiment, the interpolation between the keyframes using a Bayesian approach with a trade-off between the likelihood of the mouth shape to belong to the extracted viseme at time t and an a-priori
10 assumption on the smoothness of the trajectory.

In the following further embodiments will be described in somewhat more detail.

The embodiment described in the following makes use of a specific 3D face model called the
15 Morphable Model (MM) and which is described in V. Blanz and T. Vetter, "A Morphable Model for the Synthesis of 3D Faces " SIGGRAPH 99, 187-194. However, any model which is capable of representing the animated visemes in a viseme space where a viseme consists of individual components representing the viseme may be used.

20 In this embodiment there will be described a face animation system that is able to generate a realistic and plausible animation of the mouth (including the teeth) for a specific person. For the mouth movements there are used statistical models of the mouth shape (i.e. one statistical model for each of the 13 visemes) that are learned from data.

25 For this, several steps are performed.

In a first step the visemes are acquired or obtained.

The visemes are obtained with a known scanning device. According to one embodiment, per
30 identity (i.e. per person), there are acquired 16 scans (the set of 13 visemes and 3 neutral scans with closed mouth as reference). In the embodiment there are scanned the visemes of 12 identities, giving in total 192 scans. Other values for the number of scans or persons may be chosen as well.

35 In a next step the face animation model is generated. The Face Animation Model (FAM) is a generative statistical model that is able to generate all possible mouth movements for all visemes. The FAM is used together with the Morphable Model (MM): the MM codes the identity and is used to compute the 3D reconstruction. The FAM codes the mouth movements

(and does not affect the identity) and is used for the face animation.

To compute the FAM, first the scans are brought into correspondence, this defines a common parametrization (i.e. a common mesh) for all scans, as shown in the 3rd row of Fig. 1, for the example. Fig. 1 shows some example scans for the viseme ao for three identities (1st, 2nd, 3rd columns resp.). The second row shows the raw scan, the third row the preprocessed data

The registered scan of person i for viseme v is denoted with x_{vi} and the corresponding neutral-free scan

$$\xi_i^v = x_i^v - x_i^{neut}$$

is computed by subtracting the neutral scan x_i^{neut} of the same individual.

The neutral-free scans for all visemes are stacked to the data matrix

$$X = [\xi_1^{ao}, \xi_2^{ao}, \dots, \xi_1^v, \dots, \xi_i^v, \dots]$$

To compute the components of the model, there is used Principle Component Analysis (PCA) on the set of example scans

$$X = VWY$$

where the components $V = [v_1, \dots, v_n]$ are (column-) orthonormal, W is diagonal and $\sigma_{iv} = w_i/\sqrt{n}$ where n is the number of scans. The resulting model which is learned from data is visualized in Fig. 2.

As mentioned earlier, the face animation model FAM is used together with the MM: each animated facial shape

$$x = \underbrace{\mu + \sum \alpha_i \sigma_i^u u_i}_{\text{Fitting}} + \underbrace{\sum \beta_i \sigma_i^v v_i}_{\text{Fitting}} = \underbrace{\mu + \xi_{id}}_{\text{Fitting}} + \xi_v$$

30

is represented as the MM mean μ and a linear combination of identity components u_i (given by the MM and the fitting) and linear combination of components for the animation (the

previously computed v_i together with the weighting σ_{iv}). Each shape is then coded by its identity coefficients α_i and its animation coefficients β_i .

5 In the next step, for the speech animation, there is obtained a statistical model that describes the variability of each of the 13 visemes which have been scanned within the Face Animation Model. To compute such models, all scans i of a specific viseme v are projected into the FAM

$$\beta_{i,v} = \text{diag}(\sigma_i^v) V^T \xi_i^v$$

10

and a multivariate Gaussian distribution is fitted to the data points again using PCA. The statistical models for each of the visemes

$$\mathcal{V} = (\text{diag}(\sigma_i), \mathbf{B}, \bar{\beta} \in R^n)$$

15

are the defined by the mean within the FAM $(\bar{\beta} \in R^n)$, a (column-) orthonormal matrix of components $\mathbf{B}$ and a weighting $\text{diag}(\sigma_i)$. Examples of such viseme models can be seen in Figure 3, which shows example viseme models for the viseme *ao* (top) and *f* (bottom).

20

In this manner a "probability distribution" of the components of the viseme space for the visemes can be obtained. The thus obtained statistical model represents the statistical variability of each of the plurality of V visemes within said face animation model based on said samples.

In the following there will be described the animation according to one embodiment.

25

In this embodiment there is used the timing information which has been obtained when acquiring the visemes together with the viseme models to render an image animation according to prerecorded sound files. It could as well be used to render a sequence of visemes derived based on written language, or just based on a concatenation of visemes which are obtained in some way.

30

The output of the speech analysis is a list of viseme $\mathcal{V}_i$ at time t_i

$$\mathcal{L} = \{\mathcal{V}_i, t_i\}$$

The animation is generated by fitting a smooth trajectory through the FAM

$$\beta(t) : R \rightarrow R^n$$

5 that defines a sequence of shapes

$$\xi(\beta(t)) = \text{diag}(\sigma_i'') \mathbf{V} \beta(t)$$

in vertex space.

10

The fitting is done using a Bayesian approach. This is illustrated in Fig. 4 which shows how for the speech animation there is fitted a smooth trajectory through the Face Animation Model. At time t a likely shape for the corresponding viseme is generated (statistical models for three different visemes are visualized). (see Figure 4).

15

The Bayesian approach involves a trade-off between the probability of the shape in keyframe f belonging to the viseme at that time and an a priori-assumption of the smoothness of the mouth movements using first and second order derivatives.

20

In other words, there is made use of an a-priori-assumption on the smoothness of the trajectory which is represented by a probabilistic representation of the smoothness. This is based on the assumption that the speed and the acceleration of a certain mouth movement (representing a transition from one shape to the other) is following a certain probability distribution such that there is some limit on the speed and acceleration.

25

The probability of the trajectory given the list of viseme

$$p(\beta|\mathcal{L}) \sim p(\mathcal{L}|\beta) \cdot p(\beta)$$

is proportional to the probability of the list of viseme given the trajectory

$$p(\mathcal{L}|\beta) = \prod p(\mathcal{V}_i|\beta(t_i)) \quad (2.1)$$

$$= \prod \dots \mathcal{L}^{-\frac{1}{2}} \{\text{diag}(\sigma_i'') \mathbf{B} \}^T (\beta(t_i) - \beta') \|^2 \quad (2.2)$$

30

(this is an application of Bayes theorem) where there is made use of the statistical viseme models obtained previously times the a-priori probability of the trajectory

$$p(\beta) = \prod p(\beta(t_i)) \quad (2.3)$$

$$= \prod \eta_1 \dots e^{-\frac{1}{2} \|\nabla \xi(\beta(t_i))\|^2} + \eta_2 \dots e^{-\frac{1}{2} \|\nabla^2 \xi(\beta(t_i))\|^2} \quad (2.4)$$

$$= \prod \eta_1 \dots e^{-\frac{1}{2} \|\xi(\beta(t_i)) - \xi(\beta(t_{i+1}))\|^2} \quad (2.5)$$

$$+ \eta_2 \dots e^{-\frac{1}{2} \|\xi(\beta(t_{i+1})) - 2\xi(\beta(t_i)) + \xi(\beta(t_{i-1}))\|^2} \quad (2.6)$$

where there are made a-priori assumptions about the smoothness of the trajectory, in particular that humans are not able to move the mouth too quickly (first derivative=speed;
5 parameter η_1) and to accelerate too rapidly (second derivative=velocity; parameter η_2).

Parameter η_1 parameter η_2 may be determined experimentally .

To maximize the probability, there is minimized the cost function

$$CF = \log p(\mathcal{V}|\beta) \cdot p(\beta) = \log p(\mathcal{V}|\beta) + \log p(\beta) \quad (2.7)$$

$$= \sum_{i=1}^M \|\text{diag}(\sigma_{i,j}) \mathbf{U}_i^T (\mathbf{s}(t_i) - \tilde{\mathbf{c}}_i)\|^2 + \quad (2.8)$$

$$\eta_1 \sum_{i=1}^{M-1} \|\mathbf{x}(\beta(t_i)) - \mathbf{x}(\beta(t_{i+1}))\|^2 + \quad (2.9)$$

$$\eta_2 \sum_{i=2}^{M-1} \|\mathbf{x}(\beta(t_{i+1})) - 2\mathbf{x}(\beta(t_i)) + \mathbf{x}(\beta(t_{i-1}))\|^2 \quad (2.10)$$

using L-BFGS-B, a standard quasi-Newton optimizer and M being the number of frames.

In the above equation 2.8 the following correction should be made: U should read W, s(ti)

should read $\mathbf{x}(\beta(t_i))$, and $\tilde{\mathbf{c}}_i$ should read $\tilde{\beta}$.

15 Instead of the above example, other optimizing functions may be used as well.

In this manner the viseme representation which gives an optimum trajectory through the viseme space can be obtained and can be used for rendering the mouth animation.

20 The final video is rendered using the combined warpfeld/3D approach as known from the art by rendering the resulting shapes.

It will be readily apparent to the skilled person that the methods, the elements, units and
25 apparatuses described in connection with embodiments of the invention may be implemented in hardware, in software, or as a combination of both. In particular it will be appreciated that the embodiments of the invention and the elements of modules described in connection therewith may be implemented by a computer program or

computer programs running on a computer or being executed by a microprocessor. Any apparatus implementing the invention may in particular take the form of a network entity such as a router, a server, a module acting in the network, or a mobile device such as a mobile phone, a smartphone, a PDA, or anything alike.

Claims

1. A computer-implemented method of animating a mouth of a face in accordance with a given sequence of visemes, said method comprising:

5 Graphically representing said face by a face model;

For each of a plurality V of possibly visemes, obtaining by measurement a plurality of I different scans or samples of said visemes;

Representing each of said plurality of I samples of each of said plurality of V different visemes by said face animation model to generate a matrix based on said scans or samples which spans up a viseme space such that a sequence of visemes can be represented through a trajectory through said viseme space;

Applying a Bayesian approach to obtain the best path through said viseme space for said given sequence of visemes,

wherein the probability of the trajectory given the list of visemes

$$15 \quad p(\beta|\mathcal{L}) \sim p(\mathcal{L}|\beta) \cdot p(\beta)$$

is proportional to the probability of the list of visemes given the trajectory

$$p(\mathcal{L}|\beta) = \prod p(\mathcal{V}_i|\beta(t_i))$$

20 2. The method of claim 1, further comprising:

Performing a principle component analysis based on said matrix to obtain the principle components for each of said plurality of visemes;

Obtaining a statistical model which represents the statistical variability of each of the plurality of V visemes within said face animation model based on said samples;

25 Representing the prior probability of a certain viseme representation by a probabilistic representation of the smoothness of the trajectory through said viseme space;

Based on said prior probability and said statistical model, applying said Bayesian approach to obtain the best path through said viseme space for said given sequence of visemes.

30

3. The method of one of claims 1 to 3, wherein

said plurality of I samples are samples of a certain visemes spoken by different subjects;
or

35 said plurality of I samples are samples of a certain viseme belonging to different words or spoken by a subject at different times.

4. The method of one of claim 1 to 4, wherein
 registered scan of person i for viseme v is denoted with x_{vi} , the corresponding
 neutral-free scan

$$\xi_i^v = x_i^v - x_i^{\text{neutral}}$$

is computed by subtracting the neutral scan x_i^{neutral} of the same individual. i, and
 the neutral-free scans for all visemes are stacked to the data matrix

5. The method of one of the preceding claims, wherein

said obtaining a statistical model that describes the variability of each of the visemes
 which have been scanned within the Face Animation Model comprises:
 projecting all scans i of a specific viseme v into the Face Animation Model by

$$\beta_{i,v} = \text{diag}(\sigma_{iv}) V^T \xi_{vi}$$

and fitting a multivariate Gaussian distribution to the data points, preferably again using
 PCA.

6. The method of one of claims 1 to 5, wherein
 the statistical models for each of the visemes

$$\mathcal{V} = (\text{diag}(\sigma_i), \mathbf{B}, \bar{\beta} \in R^n)$$

are the defined by the mean within the FAM $(\bar{\beta} \in R^n)$, a (column-) orthonormal
 matrix of components B and a weighting $\text{diag}(\sigma_i)$.

7. The method of one of claims 1 to 6, wherein

for animating a list of viseme $\mathcal{V}_i$ at time t_i

30

$$\mathcal{L} = (\mathcal{V}_i, t_i)$$

the animation is generated by fitting a smooth trajectory through the
 Face Animation Model

$$\beta(t) : R \rightarrow R^n$$

that defines a sequence of shapes

$$\xi(\beta(t)) = \text{diag}(\sigma_i^v) \mathbf{V} \beta(t)$$

in vertex space.

10

8. The method of one of claims 1 to 7, wherein

$$\prod p(\mathcal{V}_i | \beta(t_i)) = \prod \dots e^{-\frac{1}{2} \|\text{diag}(\sigma_i)^j \mathbf{B}^j \mathbf{T}^T (\beta(t_i) - \bar{\beta}^j)\|^2}$$

9. The method one of claims 1 to 8, wherein

$$\begin{aligned} p(\beta) &= \prod p(\beta(t_i)) \\ &= \prod \eta_1 \dots e^{-\frac{1}{2} \|\nabla \xi(\beta(t_i))\|^2} + \eta_2 \dots e^{-\frac{1}{2} \|\nabla^2 \xi(\beta(t_i))\|^2} \end{aligned}$$

15

10. The method of one of claims 5 to 9, wherein

maximize the probability, there is minimized the cost function

$$CF = \log p(\mathcal{V} | \beta) \cdot p(\beta) = \log p(\mathcal{V} | \beta) + \log p(\beta)$$

20

11. An apparatus for animating a mouth of a face in accordance with a given sequence of visemes, said apparatus comprising:

a module for graphically representing said face by a face model;

a module for, for each of a plurality V of possibly visemes, obtaining by measurement a plurality of I different scans or samples of said visemes;

25

a module for representing each of said plurality of I samples of each of said plurality of V different visemes by said face animation model to generate a matrix based on said scans or samples which spans up a viseme space such that a sequence of visemes can be represented through a trajectory through said viseme space;

30

a module for applying a Bayesian approach to obtain the best path through said viseme space for said given sequence of visemes.

12. The apparatus of claim 11, further comprising:
a module for carrying out a method according to one of claims 2 to 10.

5 13. A computer program comprising computer executable code which when being
executed on a computer enables said computer to carry out a method according to
one of claims 1 to 10.

ABSTRACT

- 5 A computer-implemented method of animating a mouth of a face in accordance with a given sequence of visemes, said method comprising:
graphically representing said face by a face model;
For each of a plurality V of possibly visemes, obtaining by measurement a plurality of I different scans or samples of said visemes;
- 10 representing each of said plurality of I samples of each of said plurality of V different visemes by said face animation model to generate a matrix based on said scans or samples which spans up a viseme space such that a sequence of visemes can be represented through a trajectory through said viseme space;
- 15 applying a Bayesian approach to obtain the best path through said viseme space for said given sequence of visemes.

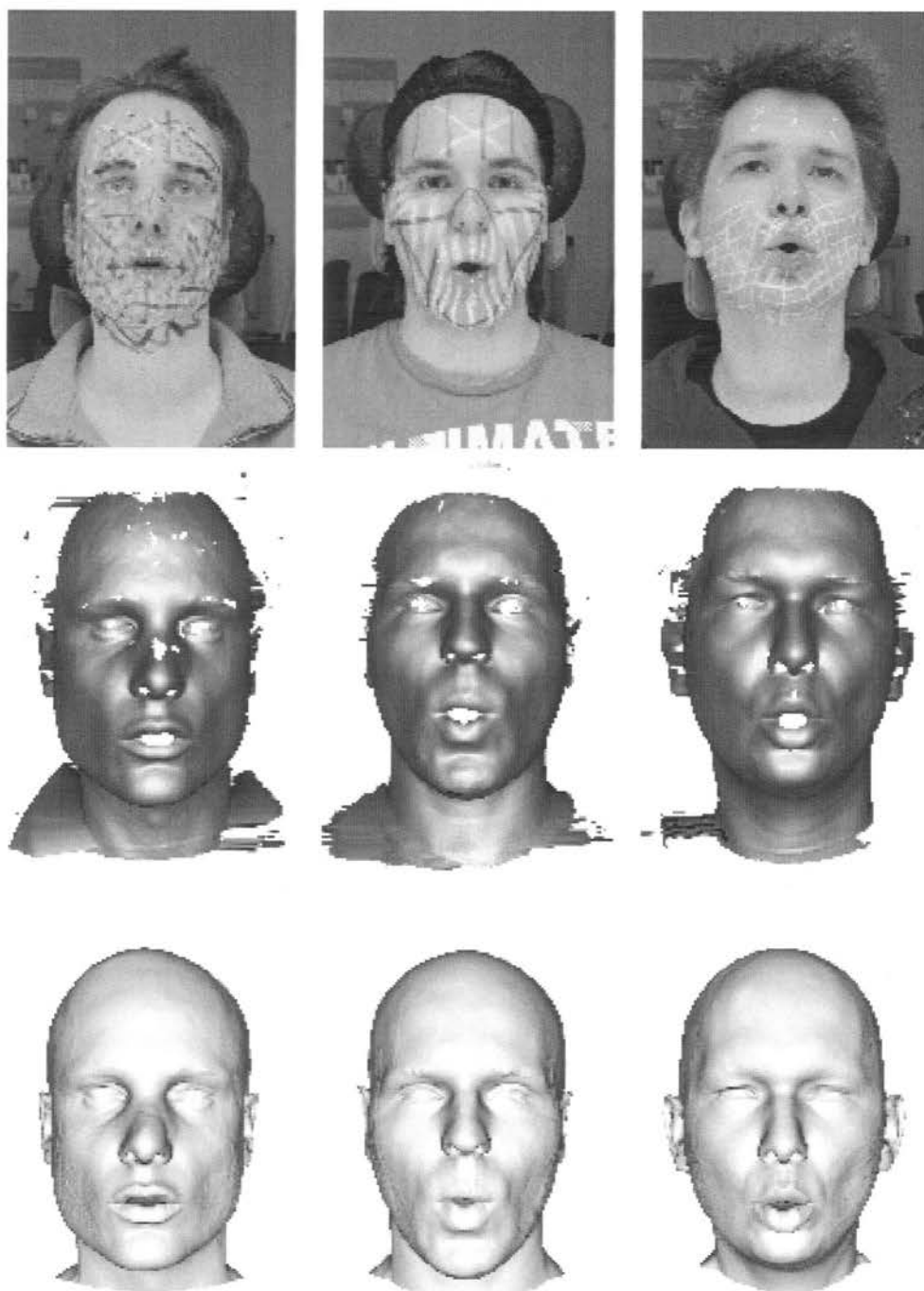
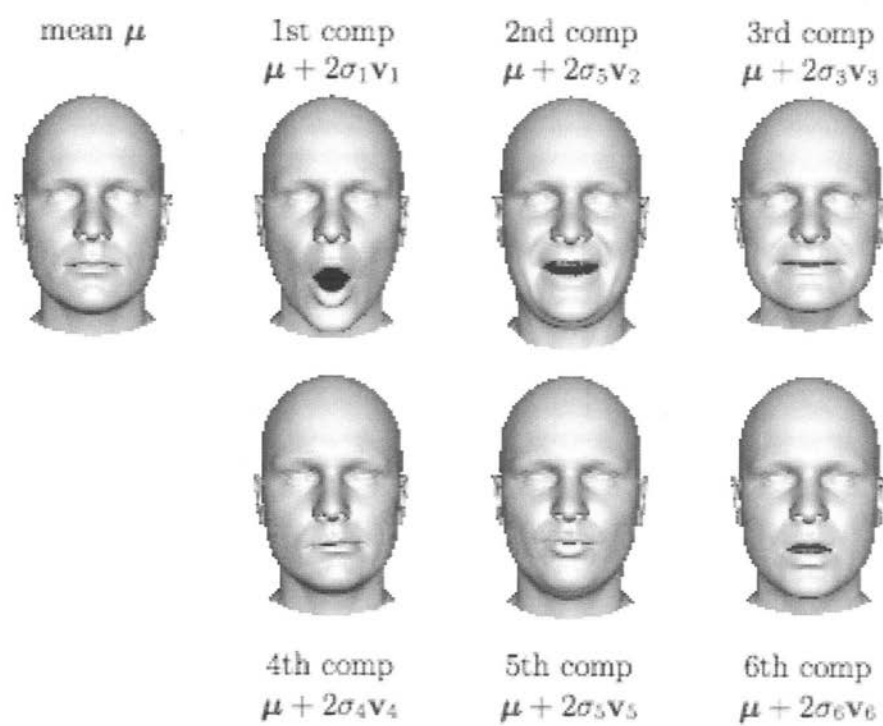
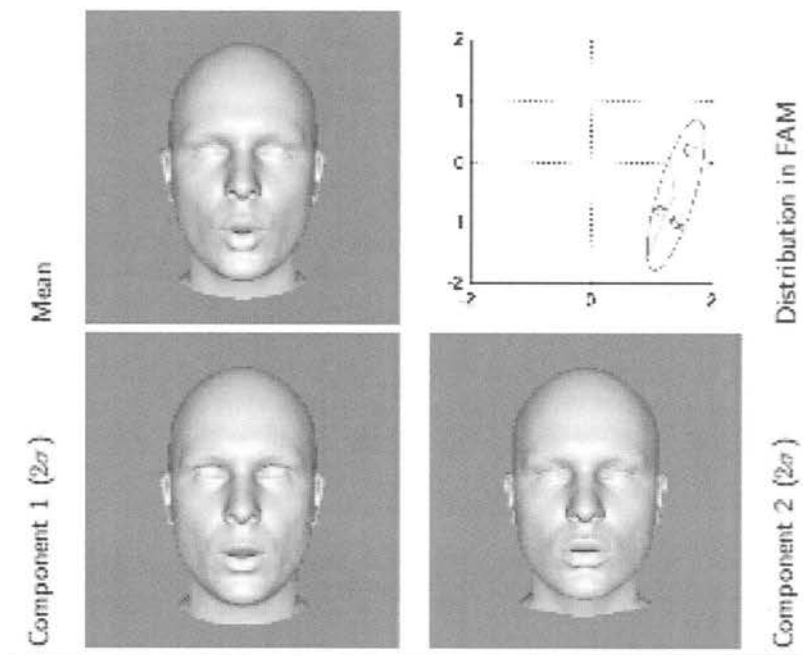
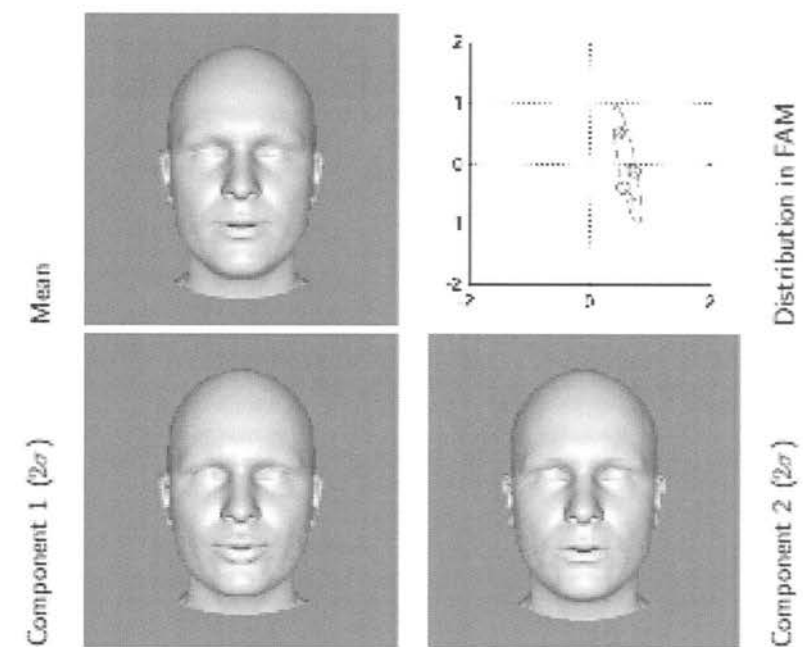
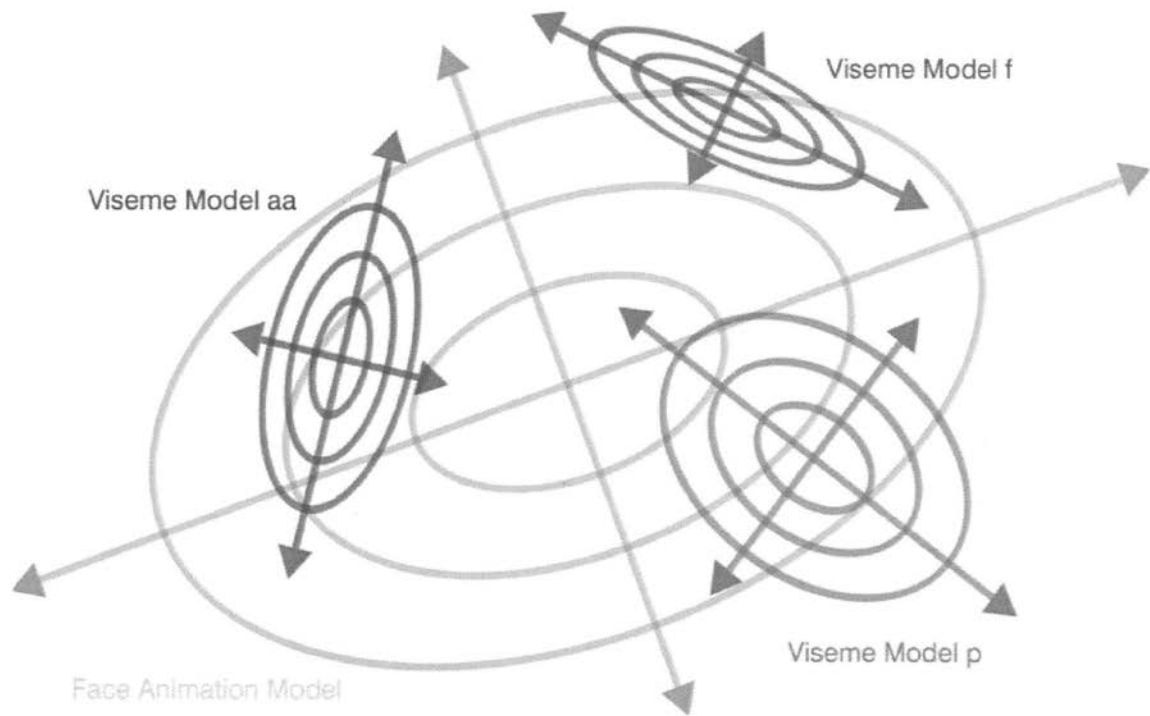


Fig. 1

**Fig. 2**

Viseme ao**Viseme f****Fig. 3**

**Fig. 4**