正面顔画像からの顔形状ディスプレイ用テクスチャ自動生成

Automatic Texture Generation for Face-Shaped Display from a Frontal Face Image

前島 謙宣† 倉立 尚明‡ Brennand PIERCE‡ ゴードン チェン‡ 森島 繁生†

Akinobu MAEJIMA† Takaaki KURATATE‡ Brennand PIERCE‡ Gordon CHENG‡ and Shigeo MORISHIMA†

† 早稲田大学 † Waseda University
‡ ミュンヘン工科大学 ‡ Technical University Munich

E-mail: †{a.maejima@kurenai, shigeo@} waseda.jp, ‡{kuratate, bren, gordon}@tum.de

1. はじめに

リアルな見た目を持ち自然な対話が可能なヒューマノイドロボットの実現は，人-ロボット間のインタラクションを円滑にするために重要であり，この目的を達成するために様々な研究が行われている[1~5].

Kuratate らは，Mask-bot と呼ばれるヒューマノイドロボットの顔を開発している (図 1). Mask-bot は， 主に 3 次元顔形状スクリーン，プロジェクタ，パンチルトユニットの 3 つから構成されており，CG 合成されたフェイシャルアニメーションを顔形状スクリーンに投影することで，自身の表情を制御することが可能で，これにより，立体視とは異なり実空間で存在感のある顔を表現することができる．また，顔のテクスチャを変更することにより，自身の顔を様々な人の顔に入れ替えることも原理的に可能である．しかしながら，実際に対象人物の顔の 3 次元形状を反映して顔の入れ替えを行うためには， 3 次元顔モデルを生成した上で，さらにそれと顔形状スクリーンとの間のキャリブレーションを逐一とらねばならず，この作業に手間と数分の時間を要するという問題が指摘されていた[1].

このような問題を解決するため本稿では，予めスクリーンとのキャリブレーションが済んでいる 3 次元顔モデルを基準モデルとし，この基準モデルに合うように正規化された，顔形状ディスプレイ用の個人顔テクスチャを正面顔画像 1 枚から自動生成する手法を提案する.

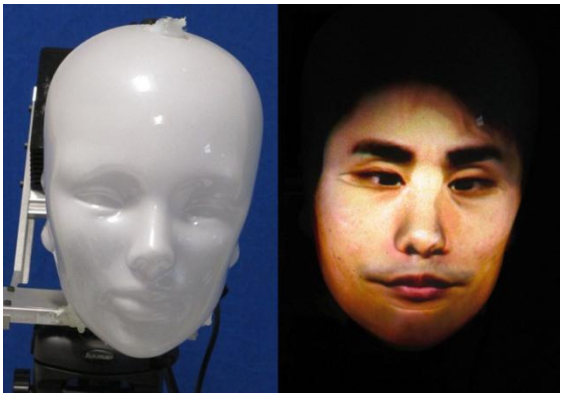

図 1. Mask-bot(左)と実際の投影の様子(右)

正面顔画像から取得困難な顔の側面部分のテクスチャは，テクスチャモーファブルモデルを利用して補完される．提案手法により，約 3 秒で Mask-bot の顔を入れ替えることが可能となり，Mask-bot のビデオ会議や認知実験への適用を容易にすることができた.また，本稿で提案するフレームワークは統一されたテクスチャ座標を持つ 3 次元顔モデル群のテクスチャ生成にも利用することができ，ゲーム制作現場での応用も可能であると考えられる.

2. 関連研究

本研究に最も類似した研究として Hayashi らは，可変形な顔形状スクリーンに正面から顔アニメーション投影することで，対象人物に近い形状と見た目を表現可能なロボットの顔を開発している[4]. スクリーンの形状は，対象人物の顔上に配置された制御点の 3 次元位置座標により決定され，そこに投影される顔アニメーションは，対象人物のビデオ映像そのものである. このため，撮影時に顔形状スクリーンと対象人物の顔の間の姿勢の対応が取れない場合，このずれによる違和感が生じる可能性がある.これに対して提案手法は，常に基準モデルに正規化された個人顔テクスチャを生成し，アニメーションを合成するため，利用者は，正面顔画像撮影の数秒の間だけカメラに対して正面を向くだけで良い．このため提案手法は利用者に対する負担がより軽いといえる.

正面顔画像から取得困難な顔側面のテクスチャを推定する代表的な手法として，Blanz らによって提案されている 3D Morphable Model に基づく方法が挙げられる[8]. 本研究においても，3D Morphable Model のテ

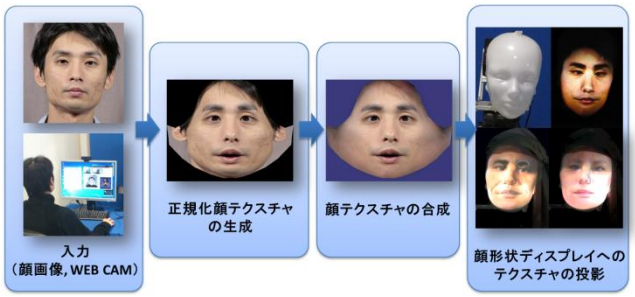

図 2 提案手法の全体像

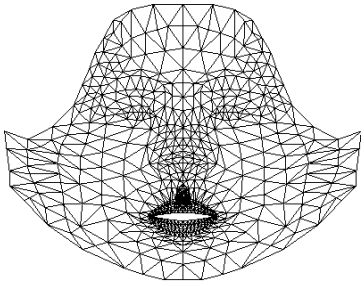

図 3. 基準モデル

クスチャ部と同様，テクスチャを既存のテクスチャの線形結合で近似し，元画像と合成するアプローチをとる．ただし，モデルフィッティング処理の高速化のため，多重解像度モデルによるフィッティングは行わず，代わりに，画像上の評価するピクセルを，テクスチャの推定結果が見た目に変わらない程度に間引き・選択する工夫をする．

なお本研究では，実際の利用場面を想定して入力画像中の顔はカメラに対してほぼ正面を向いていることを仮定する．より大きな顔の姿勢変動への対応は，例えば前処理として，顔画像から検出される特徴点に対して標準的な 3 次元顔モデルを剛体変換により整合し，その後，整合された 3 次元顔モデルを正面向きに揃えレンダリングされた，姿勢変動のキャンセルされた画像を入力とすることにより実現可能であると考える．

3. 提案手法の全体像

本研究では，予め顔形状スクリーンとのキャリブレーションが済んでいる 3 次元顔モデルを基準モデルとし，基準モデルの形状に合うように正規化された顔形状ディスプレイ用の個人顔テクスチャを正面顔画像から生成する．提案手法の全体像を図 2 に示す．

まず正面顔画像から，基準モデル(図 3)の形状に合ったテクスチャを生成する(本稿ではこれを正規化顔テクスチャと呼ぶ)．次に，テクスチャモーファブルモデルを用いて，正規化顔テクスチャに対するテクスチャ推定を行い、最後に正規化顔テクスチャと得

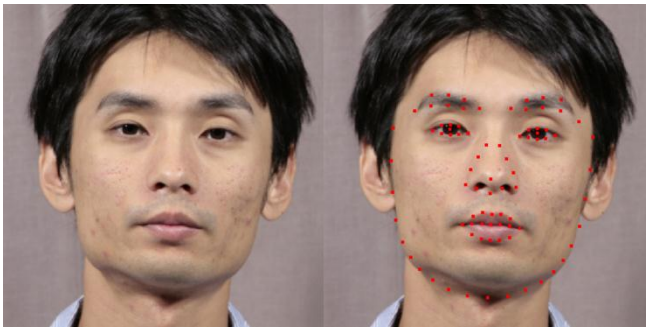

図 4 入力顔画像(左)と特徴点検出結果(右)

られた結果をリニアブレンディングすることにより最終的な顔形状ディスプレイ用のテクスチャが生成される．生成されたテクスチャは，生成後直ちに顔形状ディスプレイに投影される．

4. 正規化顔テクスチャの生成

正規化顔テクスチャは以下の手順で生成される．まず，入力の正面顔画像から，Zhang らの手法[2]により 84 点の特徴点を検出する(図 4)，次に，検出された特徴点と基準モデル上の対応する頂点の組を用いて，Radial Basis Functions により基準モデルを画像に整合し，入力画像に対応するテクスチャ座標を取得する．

$$f_i(\mathbf{x}) = \sum_{j=1}^{N} w_j^i \phi(\mathbf{x} - \mathbf{x}_j) \quad \ldots(1)$$

ここで，$f_i(\mathbf{x})$ は，$i = x, y$ 成分に対する Radial Basis Functions を表し，これが特徴点の座標と求めたいテクスチャ座標に相当する．N は特徴点とそれに対応する基準モデル上の頂点の数であり，$\mathbf{x} = (x, y)'$ は，基準モデル上の任意の頂点，$\mathbf{x}_j = (x_j, y_j)'$ は，特徴点に対応する j 番目の頂点の（基底関数の中心の）xy 座標を表す．また，w_j^i は，$i = x, y$ 成分に対する j 番目の基底関数に対する重みであり，ϕ は基底関数である．本研究では，特徴点が疎に分布している領域に対しては小さな変形を，密に分布している領域に対してはその分大きな変形を与えることのできる Hardy multi-quadrics[7]を基底関数として用いる．

$$\phi(\mathbf{x} - \mathbf{x}_j) = \sqrt{\left\|\mathbf{x} - \mathbf{x}_j\right\|^2 + s_j^2} \quad \ldots(2)$$

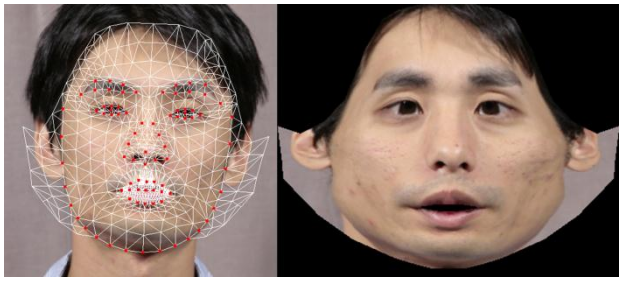

図 5 基準モデルの整合結果(左)と
正規化顔テクスチャ(右)

ここで s_j は，点 $\mathbf{x}_j$ に最も近い基準モデル上の点との距離で以下のように表される．

$$s_j = \min_{i \neq j} \left\| \mathbf{x}_i - \mathbf{x}_j \right\| \quad \ldots(3)$$

式(1)を解いて得られた正面顔画像に対するテクスチャ座標を用いて，基準モデルに正面顔画像をテクスチャマッピングし，レンダリングすることにより正規化顔テクスチャが生成される．生成された正規化顔テクスチャの例を図 5 に示す．

5. 顔テクスチャの合成

正面顔画像から取得困難な顔側面のテクスチャを，テクスチャモーファブルモデルを用いて推定し，最終的な顔テクスチャを合成する．テクスチャモーファブルモデルは，複数テクスチャの線形結合により任意のテクスチャを表現可能なモデルで，予め正規化顔モデルと同じ形状に統一された 130 枚のテクスチャに対して主成分分析を行うことにより構成される．図 6 にテクスチャモーファブルモデルの平均と第 1〜3 基底ベクトルの＋3σ 成分を示す．

4 章で述べた正規化顔テクスチャが与えられたとき，個人顔に対するテクスチャの推定は，式(4)を解くことにより得ることができる．

$$\arg\min_{\mathbf{a}} \frac{1}{2} \sum_{i=1}^{N} \left\| \mathbf{W}_{\kappa_i} (\mathbf{d}_{\kappa_i} - \mathbf{U}'_{\kappa_i} \mathbf{a}) \right\|_2^2 \quad \ldots(4)$$

ここで，$'$ は転置を表し，$\mathbf{U}$ はテクスチャモーファブルモデルを構成する基底ベクトルの集合，$\mathbf{a}$ はそれに対するパラメータ，$\mathbf{d}$ は正規化顔テクスチャの画素のベクトル形式 $\mathbf{x}$ とテクスチャモーファブルモデルの平均ベクトル $\bar{\mathbf{x}}$ との差分，$\mathbf{W} = \mathrm{diag}(w_1, \ldots, w_n)$ は，テクスチャ基底に対する重み，すなわち推定における各画素の寄与度を表す．また，κ_i は最適化の際に評価する画素のインデックスを表し，N はその個数である．

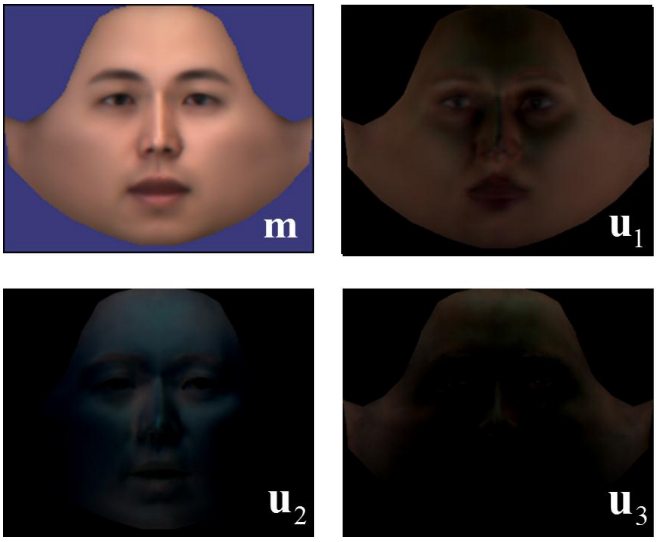

図 6 テクスチャモーファブルモデル
平均と第 1〜3 基底ベクトル

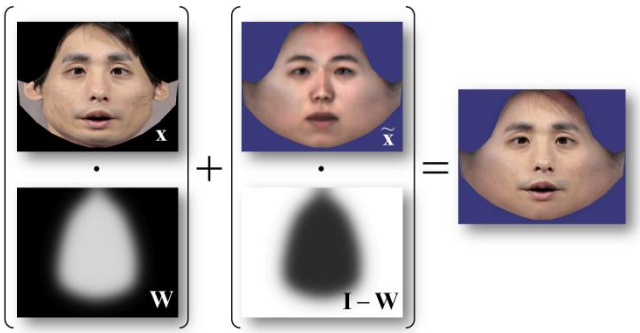

図 7 正規化顔テクスチャと推定結果の
ブレンディングによるテクスチャの合成

本稿では，入力画像に対して縦横 k 画素間隔の一様格子点上の点をサンプリングすることにより評価すべき画素のインデックスを決定している．推定結果のテクスチャの品質を考慮しつつ，k の値を大きくし評価点を間引くことで処理の高速化を実現する．なお，最適な k の値に対する考察については次章で述べる．

式(1)を非線形最適化手法である Levenberg-Marqardt 法を用いて解き，得られた $\mathbf{a}$ に基づき $\tilde{\mathbf{x}} = \mathbf{U}\mathbf{a} + \bar{\mathbf{x}}$ を計算することによりテクスチャの推定結果を得る．

最後に得られた推定結果と正規化顔テクスチャを重み行列 $\mathbf{W}$ に従ってブレンディングすることより，最終的な顔形状ディスプレイ用のテクスチャを得ることができる(図 7)．

$$\hat{\mathbf{x}} = \mathbf{W}\mathbf{x} + (\mathbf{I} - \mathbf{W})\tilde{\mathbf{x}} \quad \ldots(5)$$

テクスチャは生成されるとネットワーク経由で直ちに顔形状ディスプレイに投影される．

6. 実験

6.1 実験 1：最適なサンプリング間隔 k の決定

テクスチャ推定高速化のための最適な評価点のサンプリング間隔 k を決定するため，$k = 2^0, 2^1, .., 2^4$ とし，テクスチャ推定を行ったときの処理時間とテクスチャの平均推定誤差を計測し，さらに生成されたテクスチャとピクセル毎の誤差を誤差マップとして目視で確認した．テクスチャの平均推定誤差は式(6)で定義される．

$$E = \frac{1}{3N}\sum_{i=1}^{N}\left\|\mathbf{T}_i^{tgt} - \mathbf{T}_i^{est}\right\| \quad ...(6)$$

ここで，$\mathbf{T}_i^{tgt}$，$\mathbf{T}_i^{est}$ は，ターゲットおよび推定結果のテクスチャの i 番目の画素の RGB 値を要素に含むベクトルを表し，N は評価点の数を表す．なお，本実験では，推定に用いるテクスチャモーファブルモデルの基底の数は 20 に固定した．実験環境は，Intel Corei7 3.4GHz 8GB RAM であり，提案手法は C#, DirectX9 により実装した．

実験結果を図 8 に示す．図中，速度比は評価点の間引きを行わない状態($k = 2^0$)と間引きを行った場合との平均処理速度の比を表している．また，誤差マップは，青色が誤差 0，赤色が誤差 10 以上(各ピクセルにおけるターゲットと推定結果の RGB チャンネルの輝度の差の和が 10 以上)を表す．図 8 から，$k = 2^3$ としたとき，テクスチャの推定結果の品質を保持しつつ，約 10 倍程度高速化できることが分かった．

6.2 実験 2：テクスチャモーファブルモデルの最適な基底数の決定

テクスチャ推定に用いるモーファブルモデルの最適な基底数を決定するため，30 人分正面顔画像からテクスチャ推定実験を行う． 実験では，テクスチャモーファブルモデルにおける基底ベクトルの数を 1, 2, 5,10, 20, 50, 100, 130 と変化させて，その時の平均推定誤差と平均処理時間を計測する．実験環境は 6.1 節と同様である．また，評価点のサンプリング間隔は，6.1 節の結果から最適と思われる値である $k = 2^3$ とした．

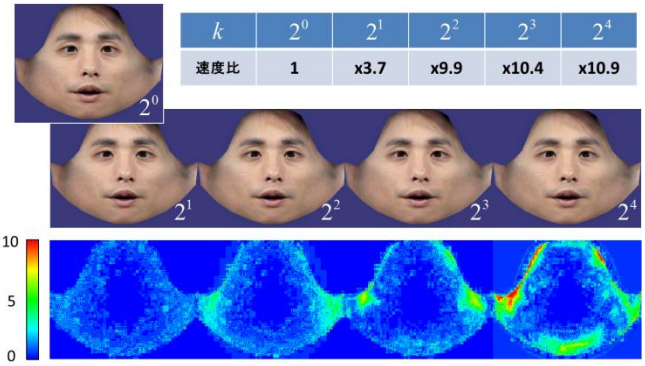

図 8 評価点のサンプリング間隔 k によるテクスチャの平均推定誤差と処理の速度比

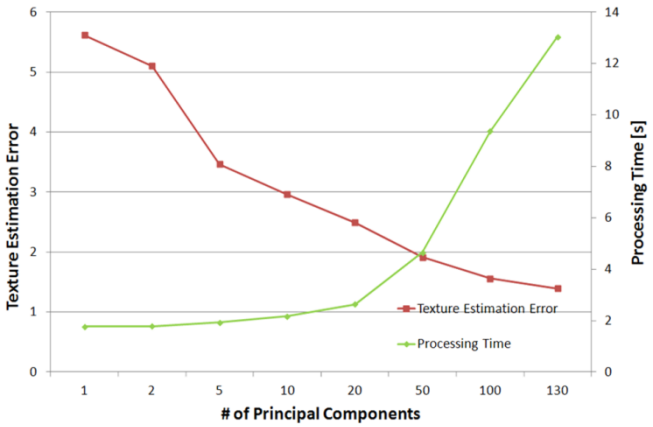

図 9 モーファブルモデルの基底数の違いによるテクスチャの平均推定誤差と平均処理時間の関係

平均推定誤差と平均処理時間の関係を図 9 に，また図 10 に入力の正面顔画像と基底数 1, 10, 20, 50, 130 のときに生成されたテクスチャ示す．図 9 における横軸はテクスチャ基底の数を，縦軸は平均推定誤差と平均処理時間を表している．

図 9 および図 10 から，テクスチャモーファブルモデルの基底数を増やせば増やすほど平均推定誤差は減少し，処理時間が指数的に増加する傾向がある．またある基底数を境にテクスチャの見えの品質が劣化していくことが見て取れる．品質劣化は高次基底に含まれるノイズに近い高調波成分をフィッティングすることに起因するものと考えられる．従って，これらの結果から，テクスチャモーファブルモデルの基底数は 20 が最適であり，このとき平均 2.6 秒でテクスチャを生成することが可能であることが分かった．

生成されたテクスチャを元に作られた実際のロボットの顔の様子を図 11 と補足動画に示す．提案手法は，Blanz らが提案する 3D Morphable Model [3]のテクスチャ推定部と比較してシンプルな実装であるが，図 11 および動画を見る限り品質は十分満足できるものであ

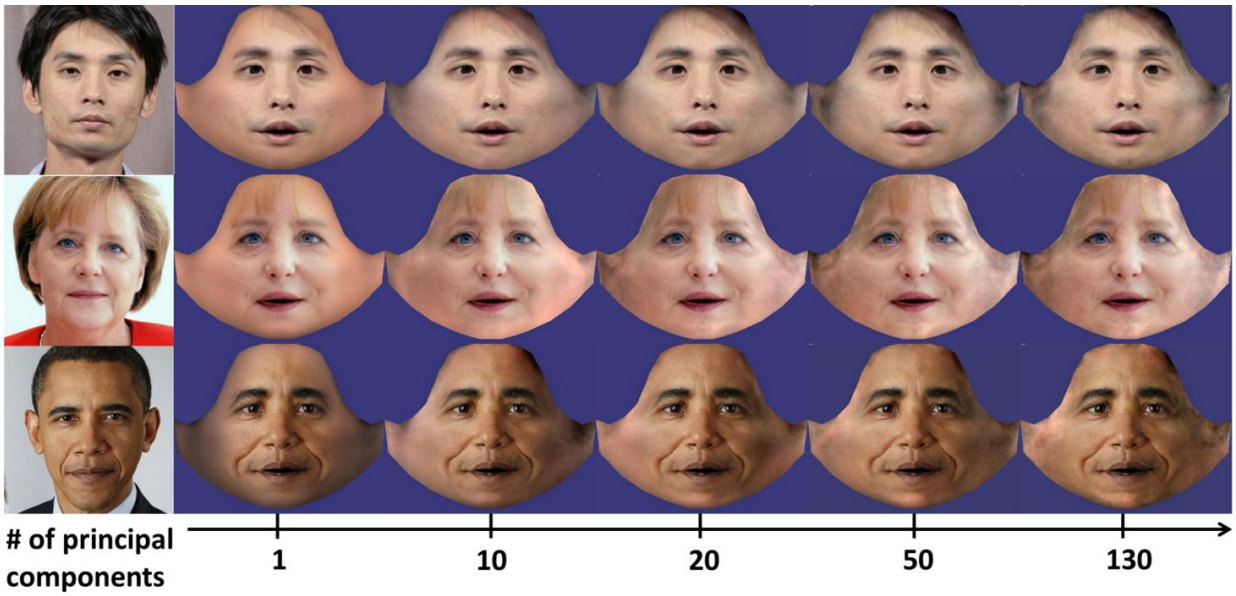

図 10. モーファブルモデルの基底数の違いによる推定結果の見えの違い

る．また，3D Morphable Model を基にした実装[10]が，形状とテクスチャの推定に平均 27.2 秒(+手作業による初期位置合わせ)を要するのに対して，提案手法は平均 2.6 秒で顔形状ディスプレイに適したテクスチャを自動生成することが可能である．このため提案手法はより実用的であるといえる．

7. おわりに

本稿では，予めキャリブレーションがとられた 3 次元顔モデルを基準モデルとし，基準モデルの形状に合うように正規化された顔形状ディスプレイ用の個人顔テクスチャを正面顔画像 1 枚から自動生成する手法を提案した．実験結果から，約 3 秒の処理時間で個人の顔テクスチャを生成可能であることが分かり，提案手法により，Mask-bot のビデオ会議や認知実験への適用を容易にすることができた．また，提案手法は，より容易にマスクを交換可能な次世代版の Mask-bot[11]への適用も容易である．

提案手法により合成されるテクスチャは，モーファブルモデルの構築に用いられるテクスチャの線形結合で表現されるため，このモデルのみで，しみ，しわ，肌理といった肌の詳細を表現することは難しい．したがって実際にゲーム制作現場での応用を考えた場合には，まずこの問題を解決する必要がある．また，本稿では，評価点を一定間隔の格子状上の点からサンプリングすることにより決定したが，より少数の，推定に有用な点のみを選択するために，Mayer と Anderson が提案しているキーポイント選択手法[9]を応用することを考えている．また，さらなる今後の課題として，提案手法により生成されたロボットの顔の個人識別に関する主観評価実験と，個人を表すのに最適な形状を持つロボットマスクの設計・開発が挙げられる．

文　献

[1] H. Ishiguro, “Understanding humans by building androids”, in SIGDAL Conference, R. Fernandez, Y. Katagiri, K. Komatani, O. Lemon, and M. Nakano, Eds. The Association for Computer Linguistics, pp. 175-175, 2010.

[2] D. Hanson, “Exploring the aesthetic range for humanoid robots”, CogSci-2006 Workshop: Toward Social Mechanisms of Android Science, 2006.

[3] C. Kroos, D. C. Herath, and Stelarc, “The articulated head pays attention”, Proc. of the 5th ACM/IEEE International Conference on Human Robot Interaction (HRI’10), pp. 357-358, 2010.

[4] K. Hayashi, Y. Onishi, K. Itoh, H. Miwa, and A. Takanishi, “Development and evaluation of face robot to express various face shape”, In Proc. of the 2006 IEEE International Conference on Robotics and Automation, ICRA 2006, pp. 481-486, 2006.

[5] T. Kuratate, B. Pierce, and G. Cheng, Mask-bot: a life size talking head animation robot for av speech and human robot communication research, In IEEE-RAS International Conference on Humanoid Robotics, pp.111-116, 2011.

[6] L. Zhang, H. AI, S. Tsukiji, and S. Lao, A fast and robust automatic face alignment system, In IEEE International Conference on Computer Vision, Demo program, 2005.

[7] J. Y. Noh, D. Fidaleo, and U. Neumann, “Animated deformations with radial basis functions”, In Proc. of the ACM symposium on Virtual reality software and technology (VRST’00), pp. 166-174, 2000.

[8] V. Blanz, and T. Vetter, A morphable model for the synthesis of 3d faces, In Proc. of the 26th annual conference on Computer graphics and interactive techniques, SIGGRAPH’99, pp.187-194, 1999.

[9] M. Mayer, and J. Anderson, “Key point subspace acceleration and soft caching”, ACM Trans. Graph. 26, 3, Article 74, 2007.

[10] FaceGenModeler, http://www.facegen.com

図 11. 提案手法により合成されたテクスチャと Mask-bot への投影結果
左: 原画像　中央: 合成されたテクスチャ　右: Mask-bot への投影結果

[11] B. Pierce, T. Kuratate, A. Maejima, S. Morishima, Y. Matsusaka, M. Durkovic, G. Cheng, “Development of an integrated multi-modal communication robotic face”, In IEEE Workshop on Advanced Robotics and its Social Impacts (ARSO) to be appeared, 2012..

謝辞

本研究はドイツ DFG cluster of excellence ’Cognition for Technical systems CoTeSys’ および, ERASMUS MUNDUS, BEAM Scholarship: L031000059 の助成を受けて実施されたものである.