論　文

身体動揺の計測による待ち状態の実写アバタ生成

宮内　翼[†]　吉村　宏紀[†]　西山　正志[†a)]　岩井　儀雄[†b)]

Generating Image-based Avatars in Wait State by Measurement of Body Sway

Tsubasa MIYAUCHI[†], Hiroki YOSHIMURA[†], Masashi NISHIYAMA[†a)], and Yoshio IWAI[†b)]

あらまし　本論文では，インタラクションシステムにおけるアバタの待ち状態に人間に近い動きを加えるために，あたかも人間と同じような身体動揺をもつ実写アバタを生成する手法について述べる．既存手法は行動状態のアバタ生成に主眼を置いており，その前後に存在する待ち状態については十分に考慮されていなかった．ユーザがアバタと円滑にインタラクションを行うためには待ち状態が重要であり，提案手法では実際の人間から計測した身体動揺を実写アバタの待ち状態で再現することを考える．提案手法は，直立姿勢を対象としカメラ映像中の人物領域から身体の各部位の振動量を計測する．得られた振動量の時間変化から特徴を抽出し，映像中の時刻をランダムに遷移することで，実写アバタの身体動揺を任意の時間長で生成する．提案手法により身体動揺を再現した実写アバタは，待ち状態において人間に近い動きをすることを主観評価で確認した．

キーワード　アバタ，身体動揺，直立姿勢

1. はじめに

近年，人間と機械とが自然にインタラクションを行うシステムに向けてアバタ [1]〜[6] が数多く開発されており，社会の様々な場面で応用が期待されている．例えば，エントランスで施設の案内をするアバタ [1]，看護実習における患者アバタ [2]，博物館で案内をするアバタ [3]，過去の体験を語るアバタ [4] などが実現されている．これらのアバタを用いることで，場所や時間の制約を受けずユーザに対してインタラクションができるようになる．本論文では，このようなインタラクションシステムの中でも，特に大型ディスプレイ上に実際の人間を映しだす実写アバタ [4]〜[6] に注目する．実写アバタは既設の大型ディスプレイを積極的に活用できる利点がある．以下では実写アバタのシステムとして，空港やチケットセンターなど特定の場所に設置された情報端末を考える．実写アバタは端末内もしくはその付近に存在し，具体的なインタラクションとしてユーザへの挨拶や端末操作の案内を想定する．

[†] 鳥取大学大学院工学研究科情報エレクトロニクス専攻，鳥取県
Graduate School of Engineering Department of Information and Electronics, Tottori University
a) E-mail: nishiyama@eecs.tottori-u.ac.jp
b) E-mail: iwai@ike.tottori-u.ac.jp

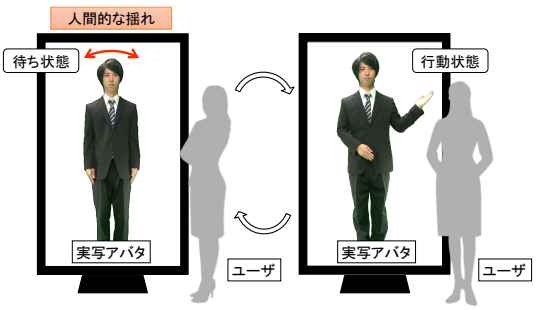


図 1　実写アバタに必要な待ち状態と行動状態．
Fig. 1　Image-based avatar system need wait state and action state.

このようなインタラクションシステムにおいて自然な実写アバタを再現するためには，ユーザとの対話などの行動状態に加えて待ち状態を取り扱う必要がある．実写アバタは人間との間で何らかの行動を常にとっている訳ではなく，行動の前後に待ち状態が存在する (図 1)．ここで述べる待ち状態には，ユーザとの対話が開始される前で相手が来ることを待つフェーズと，ユーザとの対話中に相手の反応を待つフェーズに分けられる．本論文では，前者の待ち状態を対象とし，アバタ 1 体とユーザ 1 名が存在する状況について議論する．対話前の待ち状態が適切に実写アバタに組み込ま

れていないと，ユーザは実写アバタに話しかけて良いかどうか判断できず，インタラクションが円滑に開始されない問題がある．例えば，待ち姿勢の静止画を表示し続けた場合や，人間の動きとは違う不自然な映像を表示した場合，ユーザはシステムが対話を待っている状態とは判断できず，システムが異常動作していると判断する恐れがある．実写アバタの既存手法 [4]〜[6] は行動状態を対象としていたものの，待ち状態を十分に考慮していない課題があった．

ここで，対話前の待ち状態として受付などでよく見かける直立姿勢について考える．直立姿勢の人間が待ち状態においてどのような動きをするかを実際に観察すると，ある位置を中心とし絶えず揺れ動いていることが分かる．これは身体動揺 [7] と呼ばれており，一部の筋肉に負担が掛からないように上手く負担を分散するよう人間は無意識の内に体を制御している．

そこで本論文では，実際の人間から身体動揺の振動量を計測しアバタで再現することで，待ち状態でも人間的な動きができる実写アバタを生成する手法を提案する．身体動揺を計測するために，カメラ映像から体の部位毎に振動量を求める．計測された振動量の時間変化から特徴を抽出し映像をランダムに遷移させることで，実写アバタの身体動揺を任意の時間長で生成する．提案手法を用いることで，実写アバタが人間的な動きに近づくことを主観評価で確認した．

以下では，2. で関連研究について紹介し，3. で身体動揺の計測手法とその結果について述べ，4. で身体動揺の再現手法と実写アバタ上での再現結果について述べる．最後に 5. でまとめる．

2. 関 連 研 究

情報端末における案内を想定したシステムにおいて，ユーザがアバタと自然にインタラクションするためには，アバタの立ち居振る舞いに対し，あたかも本当の人間であるかのようにユーザが感じることが望ましい．実際の案内者に近いと感じさせるアバタを実現するためには，文献 [8] でも述べられているように，インタラクションシステムの全体構成として，ユーザの状況やユーザとの会話を理解する要素技術に加えて，アバタの画像と音声を自然に生成しユーザに伝達する要素技術が必要である．特に，アバタの画像生成は見た目に直結する重要な要素技術であり，身体や顔の動きについて近年盛んに研究が進められている．既存手法 [1]〜[3] では，コンピュータグラフィクスのキャラクタにユーザの状況に応じた情報を与えることでアバタ画像を生成している．一方，実際の人物を撮影した映像を用いて実写アバタを生成する手法 [4]〜[6] が提案されている．一般的にコンピュータグラフィクスに比べて実写の方がアバタの見た目は本物の人間に近づくことが多い．ただし，実写アバタはカメラで撮影した映像しか動きを再現できない課題が存在する．この課題を解くことを目指し，撮影した映像から新たな人物画像を生成する手法 [9]〜[13] が提案されている．文献 [9], [10] では会話時の顔画像の生成について述べられており，文献 [11] では会話時の人物の全身画像の生成について述べられている．さらに文献 [12] では複数人による同時動作を編集する手法が提案されており，文献 [13] では複数の動作を時間方向に滑らかにつなぎ合わせる手法が提案されている．しかし，これらの既存手法は，行動状態の中で動きのある人物画像を取り扱っているが，待ち状態の中で微動している人物画像については十分に考慮されていなかった．文献 [4] では，あるインタラクション映像と別のインタラクション映像をつなぐため，映像の終了フレームと開始フレームとの間をモーフィングで補完する手法が提案されている．この手法のように単純なモーフィングでは，人間の身体動揺を忠実に再現しているとは言えなかった．また文献 [14] では，モーションキャプチャした被写体の関節位置の動きを用いることで，CG アバタの待ち状態を制御する手法が提案されている．ただしこの手法は関節のみを対象としているため，全身の映像を用いる実写アバタにはそのまま適用できない問題があった．提案手法では，待ち状態の人間で必ず発生する身体動揺を被写体の映像から解析し，実写アバタにおいて忠実に再現する課題に取り組む．

3. 身体動揺の計測

3.1 実写アバタで要求される身体動揺

本論文では，エントランスでの受付など周囲からの目がある状況において，直立姿勢の人間に生じる身体の揺れを取り扱う．この身体動揺を実写アバタで再現するためには，カメラ映像中でどのように人間の身体が揺れているかを計測することが必要となる．さらに，身体の部位毎に揺れ方が異なるのかどうかも合わせて確認する必要がある．実写アバタの待ち状態の映像は，直立姿勢の被写体が正面から撮影されることを想定している．ただし身体の揺れは微小であるため，カメラのフレームレートや画素数，カメラから被写体

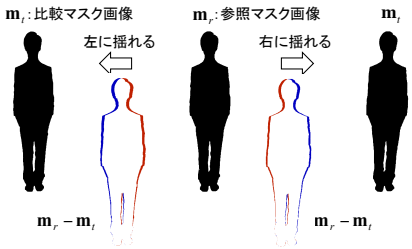


図 2　各時刻におけるマスク画像の比較．赤色は人物から背景に変化した領域を表し青色はその逆を表す．
Fig. 2　Comparing body mask images on each time.

までの距離など撮影条件の影響，および，計測手法の誤差の影響で十分に観測されない可能性もあると考えられる．本論文では，実写アバタの被写体を撮影する環境において，提案手法で本当に身体動揺が計測されるかどうかを検証する．

まず身体動揺を計測する手法について議論する．一般的に重心動揺計 [15] を用いることが多いが，この装置は足元の圧力を計測するため，体の部位毎の揺れを扱うことはできない．一方，加速度センサや角速度センサを身体の各部位に装着して計測する手法 [16], [17] が提案されている．これらの手法はカメラとセンサの間で時間同期が難しく，体に取り付けたセンサが見え隠れする問題がある．近年では複数台カメラを用いて重心位置を計測する手法 [18] や距離センサを用いて人間の関節位置から重心位置を計測する手法 [19] も提案されている．ただし，これらの手法は重心位置のみを取り扱っており部位毎の変化までは十分に検討されていなかった．本論文ではカメラ映像のみを採用し，部位毎の揺れを身体と非接触で新たに計測する．以下では，類似する揺れ方が繰り返し表れると仮定した上で設計した計測手法を 3.2 で述べ，その性能を 3.3 で確認する．

3.2　計測手法

身体動揺を計測するために人物領域の画素値を 1 とし背景領域の画素値を 0 とするマスク画像を用いる．身体動揺の中心となる参照時刻 r が与えられると，カメラ映像から参照マスク画像 $\mathbf{m}_r$ を生成し，時刻 $t \in 1, ..., N$ における比較マスク画像 $\mathbf{m}_t$ との間で，図 2 のように時間方向の差分を求める．提案手法における身体動揺の振動量 [画素] は式 (1) で計算される．

$$d_i = \sum_{x \in \mathrm{parts}(i)} (\mathbf{m}_r(x) - \mathbf{m}_t(x)) \qquad (1)$$

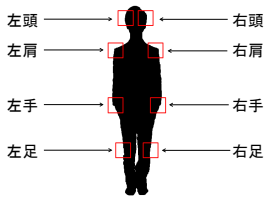


図 3　身体動揺を計測する部位領域の例．
Fig. 3　Body part regions for measuring body sway.

ただし，$\mathbf{m}_r(x), \mathbf{m}_t(x)$ は x で指定された位置の画素値を表し，$\mathrm{parts}(i)$ は身体部位 i で指定された領域を表す．これらの領域の大きさと位置は時間方向に変化させず固定とし，例えば図 3 のように設定する．指定された領域において，人物から背景に変化した画素数が多ければ d_i は正の値をとり，背景から人物に変化した画素数が多ければ d_i は負の値をとる．なお，参照マスク画像の参照時刻 r は，部位毎の振動量の合計からアルゴリズム 1 で求まる．$\tilde{d}_i$ は，ある時刻 $\tilde{r}$ を仮の参照時刻とした際の振動量を表す．

Algorithm 1 参照マスク画像の参照時刻 r の決定

```
for r̃ = 1 to N do
    D_r̃ ← 0
    for t = 1 to N do
        compute d̃_i using m_r̃, m_t
        D_r̃ ← D_r̃ + Σ|d̃_i| for all body parts
    end for
end for
r ← arg min D_r̃
```

3.3　計測性能の評価

3.3.1　振動量の計測結果

提案手法の計測性能を評価するために実験を行った．被写体 5 名 (平均年齢 21.4 ± 0.5 歳) に対し 180 秒間の直立姿勢を撮影した．被写体に撮影中はカメラへ視線を向け両足の踵をつけた姿勢を維持するよう指示した．カメラのフレームレートは 30 フレーム毎秒で解像度は 1920 × 1080 画素とした．カメラは床面から 90 センチメートルの高さに設置し，カメラと被写体の距離は 200 センチメートルとした．グリーンバックを用いた背景差分によりマスク画像を生成した．人物領域の外接矩形の大きさは約 300 × 950 画素であった．

被写体毎の振動量 d_i の標準偏差を表 1 に示す．身体部位は人手で与えた図 3 の頭，肩，手，足とし，各領域の大きさは 70 × 70 画素とした．実験結果より，被写体毎に振動量の標準偏差は異なるため，振動量に

表 1　身体動揺の振動量 d_i の標準偏差 [画素].
Table 1　Standard deviations of vibration amounts

被写体	左頭	右頭	左肩	右肩	左手	右手	左足	右足
A	483	489	328	372	284	301	117	146
B	214	185	176	123	127	88	46	48
C	306	312	253	233	216	145	83	87
D	364	434	243	339	174	184	89	98
E	321	330	258	270	204	208	91	88
平均	338	350	252	267	201	185	85	93

は個人差が存在することが分かった．ただし，部位間の大小関係は被写体間で共通であったため，開眼時にカメラに視線を向けた直立姿勢において，振動量は身体の上部位ほど大きく身体の下部位ほど小さいことが確認された．なお，今回の実験では振動量 d_i が約 380 画素ほど変化すると，身体部位が実世界で約 1 センチメートル移動したことに相当した．

次に，身体の各部位における振動量 d_i の時間変化を図 4 に示す．図中の波形は被写体間で大きく違うため，振動量の時間変化にも個人差が存在することが分かった．ただし，部位毎の時間変化に着目すると，頭が左方向に動けば肩，手，足も共に同じ方向に動いており，左側の部位が動けば右側の部位が逆の方向に動いていた．他の被写体でも同様の結果がみられた．以上の結果より，振動量の時間変化の傾向は部位間でほぼ等しいことが確認された．

さらに振動量の時間変化の特徴について検証した．ここでは被写体 A の左頭の振動量について特徴を図 5 に示す．振動量がほぼ 0 の参照時刻が時間方向に繰り返し出現しており，それらの時刻の間に存在する波形は多峰性の弧を描いていた．参照時刻は 180 秒間の中で 25 個存在し，10 秒のような間隔が短い区間でも 4 個存在していた．このように短時間の計測でも揺れの特徴が表れることが分かった．同じような特徴が他の被写体でも表れていた．以上より，振動量の時間変化の特徴として，参照時刻は映像中で複数回出現し，それら参照時刻の間における映像には身体の揺れが存在することを，実写アバタの被写体を撮影する環境において確認した．

3.3.2　計測誤差の検証

提案手法は背景差分による人物領域の抽出精度に依存しており，計測した振動量が抽出誤差に埋もれていないかを確認した．提案手法はグリーンバックを用いたが，相互反射の影響でグリーン成分が衣服領域に混じる場合や，人の影でグリーンバック領域の色が衣服の色に近づく場合が見られた．このように人物と背景の境目が曖昧な領域が存在するため以下の検証を行った．この実験では，背景差分から求めたマスク画像を用いた場合と人手で抽出したマスク画像を用いた場合とを比較した．人手による抽出は，グリーンバックと人物の境界を目視で確認しながら人物領域を設定した．被写体 5 名の参照マスク画像と，映像からランダムに選択した 15 枚の比較マスク画像を用いた．振動量の絶対値 $|d_i|$ について平均を比較した結果を図 6 に示す．この結果より，人物領域の抽出誤差の影響は十分に小さいことが分かった．これにより，3.3.1 の実験結果は人物領域の抽出誤差にそれほど影響されないと言える．

次に，距離センサで獲得した関節位置を用いて揺れの大きさを計測する場合と比較した．ここでは既存手法 [19] でも用いられている Microsoft Kinect v2 で獲得した関節位置とカラー画像を利用した．Kinect から出力される関節位置と提案手法の身体部位の位置は異なるため，この実験では以下で述べる重心を評価に用いた．Kinect を用いる手法では，同時刻のカラー画像上に射影した二次元の関節位置から重心を求めた．なお Kinect が出力する全ての関節位置を重心計算に用いた．提案手法では，Kinect のカラー画像から抽出したマスク画像に含まれる人物領域の全身を用いて重心を求めた．さらに，Kinect のカラー画像から人手で抽出した人物領域の重心を求めた．評価指標として，一定時間が経過した後に重心がどれだけ動いたかを表す移動量を用いた．Kinect で撮影した映像において，第 1 時刻をランダムに設定し，その時刻から $(0, 80)$ 秒の区間で第 2 時刻をランダムに設定した．第 1 時刻と第 2 時刻のペアを 15 組準備した．第 1 と第 2 の時刻の間の平均移動量を計算した結果，人手で抽出した場合は 3.4 画素であったのに対し，提案手法は 3.5 画素，Kinect を用いた手法は 4.7 画素であった．提案手法は，Kinect を用いた手法と比べて人手で抽出した場合に移動量が近いことが分かった．この原因として，Kinect から出力される足や手の関節位置は誤差が大きかったことが考えられる．以上より，身体の微小な動きを計測する場合，Kinect から出力される関節位置を単純に適用できるとは限らないことが分かった．

4.　実写アバタにおける身体動揺の再現

4.1　再現手法の方針

実写アバタにおいて身体動揺をどのように再現するかについて述べる．ここでは，身体動揺を計測した被

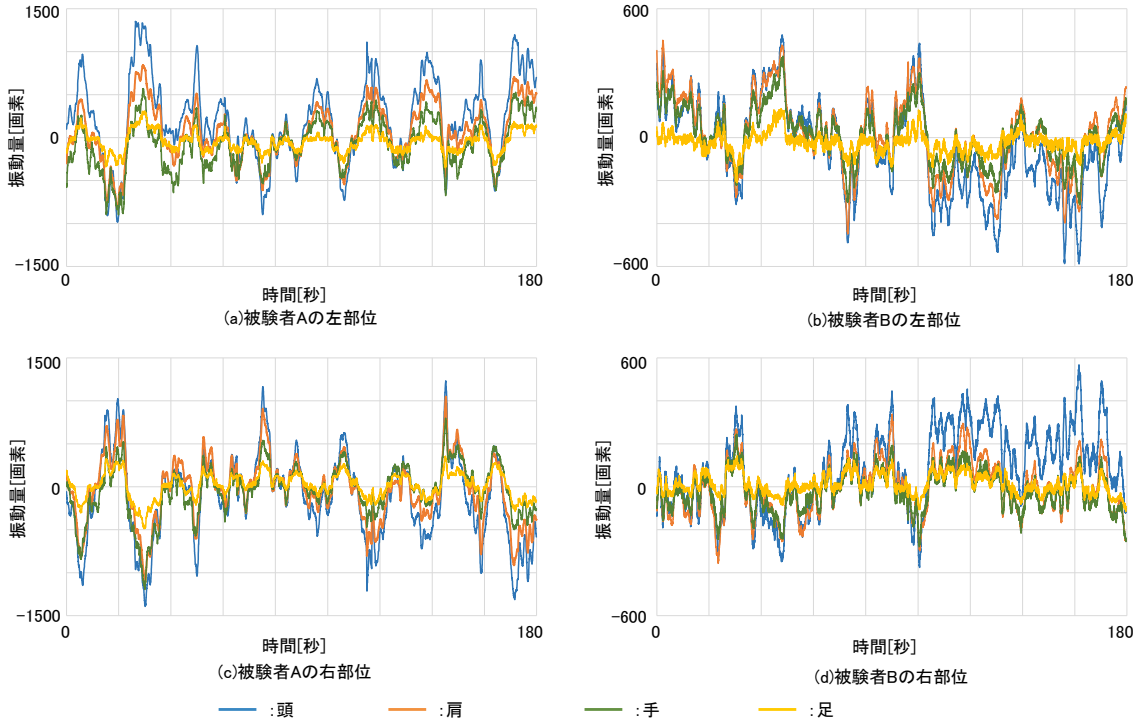


図 4　身体の各部位の振動量 d_i の時間変化.
Fig. 4　Measuring time change of movement of body sway for each body part.

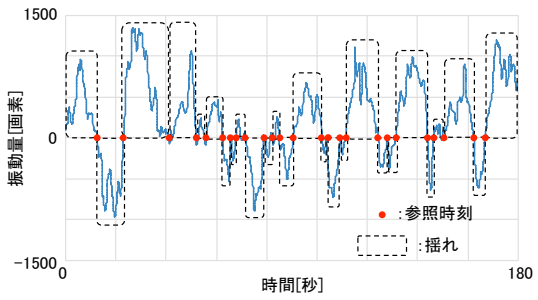


図 5　振動量の時間変化の特徴.
Fig. 5　Features of time change of vibration amounts of body sway.

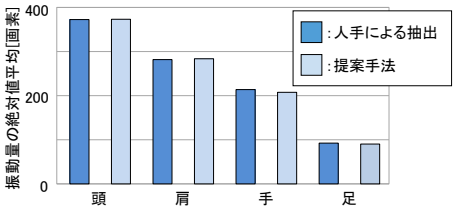


図 6　背景差分による振動量の計測性能の比較.
Fig. 6　Performance of measuring vibration amounts using a background subtraction method.

写体そのものを実写アバタとし，本人自身の揺れを任意の時間長で再現する場合を考える．この場合，図 5 で確認した身体動揺の特徴について，参照時刻が複数出現することを明示的に利用でき，参照時刻の間の揺れを暗に利用することができる．提案手法では，実写アバタのモデルを撮影した映像から身体動揺の参照時刻を複数抽出し，それらの時刻をランダムに遷移しながら映像を再生する．その流れを図 7 に示す．これにより遷移時の体の揺れを滑らかに繋げ，遷移時以外は揺れをそのまま再生することで，提案手法は身体動揺を任意の時間長で生成することができる．

ただし，ここまでの身体動揺の計測の議論では順方向の時系列しか取り扱っていないため，ランダム遷移により時刻が不連続に変化する場合は，再現時に新たな問題が発生する．ランダム遷移の参照時刻を抽出するためにマスク画像のみを用いた場合，人物領域の形状は似ているが，髪や衣服や手のずれなど身体前面のテクスチャは似ていない時刻が選ばれる可能性があ

る．このような参照時刻で遷移すると，映像中に不連続な変化が含まれるためユーザが違和感を覚える．自然な身体動揺の映像を生成するためには，形状に加えて全身の見え方も考慮する必要がある．また，ユーザは顔の見え方の変化にも敏感である．例えば目や口など顔に含まれる部品が時間方向に不連続に変化し，急に閉じたり開いたりすると違和感を覚える．このため提案手法は，参照時刻を抽出する際に顔の見え方変化も考慮に加える．以下では再現手法の詳細について述べる．

4.2 形状と見え方の類似度の算出

身体動揺を再現するために，提案手法は人物領域の形状，人物領域の見え方および顔の見え方の類似度を各時刻で求める．時刻 t における人物領域の形状を表すマスク画像を $\mathbf{m}_t$，人物領域の見え方を表すカラー画像を $\mathbf{a}_t$，顔の見え方を表す顔画像を $\mathbf{f}_t$ とする．マスク画像 $\mathbf{m}_t$ は身体動揺の計測で述べた手法と同様に背景差分から算出する．カラー画像 $\mathbf{a}_t$ はカメラ映像の各フレームとする．顔画像 $\mathbf{f}_t$ は顔追跡手法 [20] で得た特徴点を利用し顔向きを正面に正規化した領域とする．初期参照時刻 r_0 が与えられたとすると，時刻 t における人物領域の形状の類似度 $s_{t,s}$ は式 (2) で求まる．

$$s_{t,s} = e^{-\lambda \sum |d_i|} \tag{2}$$

ここで，λ は定数とし，d_i は $\mathbf{m}_{r_0}, \mathbf{m}_t$ から算出した振動量とする．この式では身体部位毎に求めた振動量の大きさの合計値を用いる．この類似度 $s_{t,s}$ を，人物領域と背景領域との入れ替わりが少ない時刻を抽出するために用いる．次に，人物領域の見え方の類似度 $s_{t,a}$ は式 (3) で求まる．

$$s_{t,a} = \mathrm{SSIM}(\mathbf{a}_{r_0}, \mathbf{a}_t) \tag{3}$$

ここで，SSIM は人の主観に近い類似度を算出する Structural SIMilarity [21] を表す．この類似度 $s_{t,a}$ を，人物領域内のテクスチャ変化が少ない時刻を抽出するために用いる．次に，顔の見え方の類似度 $s_{t,f}$ は式 (4) で求まる．

$$s_{t,f} = \mathrm{FaceSimilarity}(\mathbf{f}_{r_0}, \mathbf{f}_t) \tag{4}$$

ここで，FaceSimilarity は顔画像のエッジから特徴量を求めコサイン類似度を算出する．この類似度 $s_{t,f}$ を，目の瞬きや口の開閉など顔の表情変化が少ない時刻を抽出するために用いる．これらの類似度を加算することで統合類似度 s_t を式 (5) で求める．

$$s_t = \alpha s_{t,s} + \beta s_{t,a} + \gamma s_{t,f} \tag{5}$$

ただし，$\alpha + \beta + \gamma = 1$ とする．各類似度の値域は $[0, 1]$ であるが，実際には偏りが存在するため α, β, γ でスケーリングを行う．また，統合する際にどの類似度を重視するかの制御も可能である．統合類似度を用いることで，映像を遷移する際に身体の変化が少なくなる時刻の集合を抽出することを狙う．なお，初期参照時刻 r_0 は，s_t を用いてアルゴリズム 2 を適用することで求まる．

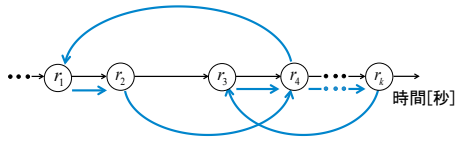


図 7 参照時刻 r_k の集合を用いた身体動揺の再現．提案手法は参照時刻をランダムに遷移し映像を再生する．図中の黒矢印は撮影した映像の時間の流れを表し，青矢印は再現した映像の時間の流れを表す．

Fig. 7 Synthesizing a video sequence of body sway using a set of reference times.

Algorithm 2 初期参照時刻 r_0 の決定．

```
for r̃_0 = 1 to N do
  S_r̃0 ← 0
  for t = 1 to N do
    compute s̃_t using m_r̃0, a_r̃0, f_r̃0, m_t, a_t, f_t
    S_r̃0 ← S_r̃0 + s̃_t
  end for
end for
r_0 ← arg max S_r̃0
```

4.3 参照時刻の集合を用いた映像遷移

自然な身体動揺の映像を再現するために，提案手法は参照時刻 r_k の集合を利用する．この集合は，初期参照時刻の身体状態に近い状態を持つ時刻をアルゴリズム 3 で抽出することで求まる．提案手法は，参照時刻の条件として，s_t の時間変化において極大点となること，かつ，s_t が閾値 T_1 より大きいことを設ける．ただし，これらの条件だけでは非常に近い時間間隔で抽出されることもあるため，閾値 T_2 以上離れた時刻を選択する条件も設ける．

次に，参照時刻の集合を用いた映像遷移の手法について述べる．抽出された時刻 r_k の集合から，ランダムに 1 つの時刻を選択し，そこから次の参照時刻になるまでのカラー画像 $\mathbf{a}_t$ をディスプレイ上で再生する

Algorithm 3 参照時刻 r_k の抽出.

```
for t = 1 to N do
  compute s_t using m_{r_0}, a_{r_0}, f_{r_0}, m_t, a_t, f_t
  if s_t is local maximum, s_t > T_1, interval > T_2 then
    add time t to set of reference times
  end if
end for
```

(アルゴリズム 4). このランダム選択を繰り返すことで身体動揺を任意の時間長で再現する. これにより, 身体動揺の振動量が異なる画像列をランダムにつなげ合わせることができ, 自然な身体動揺が実写アバタで再現される.

Algorithm 4 身体動揺の生成.

```
while true do
  select r_k randomly from set of reference times
  for t = r_k to r_{k+1} do
    display color image a_t
  end for
end while
```

4.4 再現手法の評価

4.4.1 主観評価の結果

提案手法により身体動揺を再現した実写アバタについて評価した. 図 8 の男女 2 体の実写アバタを用いた. どちらの実写アバタも直立姿勢としたが, 女性の実写アバタは手を前で組むこととした. 実写アバタのモデルとなる人物に直立姿勢を維持させ 30 秒間の映像を撮影した. この映像の先頭から 10 秒間を参照時刻の抽出対象とし, 新たな 30 秒間のアバタ映像を生成した. 抽出された参照時刻の個数は両アバタともに 4 個であった. 提案手法のパラメータは $\alpha = 0.2, \beta = 0.7, \gamma = 0.1, \lambda = 1.0 \times 10^{-5}, T_1 = 0.97, T_2 = 1$ とした. 類似度 $s_{t,s}$ を算出する際の部位領域は, 男性アバタの場合は図 3 とし, 女性アバタの場合は図 9 とした. 比較のために次の 4 つの手法を用いて主観評価を行った.

V1 : (理想) 撮影した 30 秒間の映像をそのまま再生.
V2 : (提案手法) 生成した 30 秒の映像を再生.
V3 : (比較手法 1) 先頭 10 秒間の映像を 3 回連続再生.
V4 : (比較手法 2) 一枚の静止画像を 30 秒間表示.

主観評価にはサーストンの一対比較 [22] を用いた. 被験者は 11 名 (男性 9 名, 女性 2 名, 平均年齢 21.6±0.8 歳) とし, 80 インチの縦置きディスプレイから 1.5 メートル離れた位置に立った. 各手法の映像をペアで ${}_4C_2 = 6$ 回だけランダムな順番で提示した. 被験者はペアの映像が順に再生された後に, どちらの映像がより人間の自然な動きに近いかを回答した. 一対比較の各映像を 30 秒間とした理由は, 順に映像を被験者が見ていくため, それ以上の長さとした場合は比較のために再生した前の映像の印象が薄れてしまい回答が曖昧になったためである.

図 8 評価に用いる実写アバタ.
Fig. 8 Image-based avatars for evaluations.

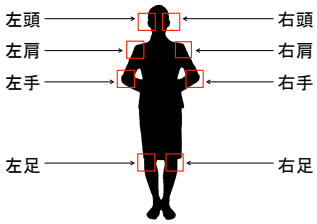


図 9 人物領域の形状の類似度を算出するために用いた部位領域 (女性の実写アバタの場合).
Fig. 9 Body parts for computing shape similarities (female image-based avatar).

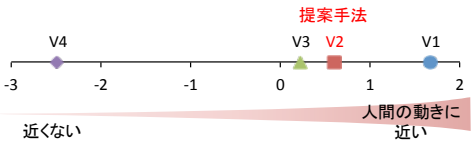


図 10 実写アバタの身体動揺に対する主観評価の結果.
Fig. 10 Subjective assessment of synthesized body sway of avatars.

主観評価の結果を図 10 に示す. 図中のスコアが大きいほどアバタの動きが人間に近いとの同意が多かったことを表す. この実験結果より, 提案手法 V2 は比較手法 V3, V4 と比べて人間の動きに近いことが分かった. 図 11 に主観評価に用いた映像の例と振動量の可視化の例を示す. この図より, V3 は連続再生の遷移時に振動量 (図中の赤色と青色の領域) が大きいため揺れに不連続が生じていることが分かる. 一対比較の後に自由記述のアンケートを取ったところ, V3 は映像の途中でアバタが急に動くことがあり不自然であっ

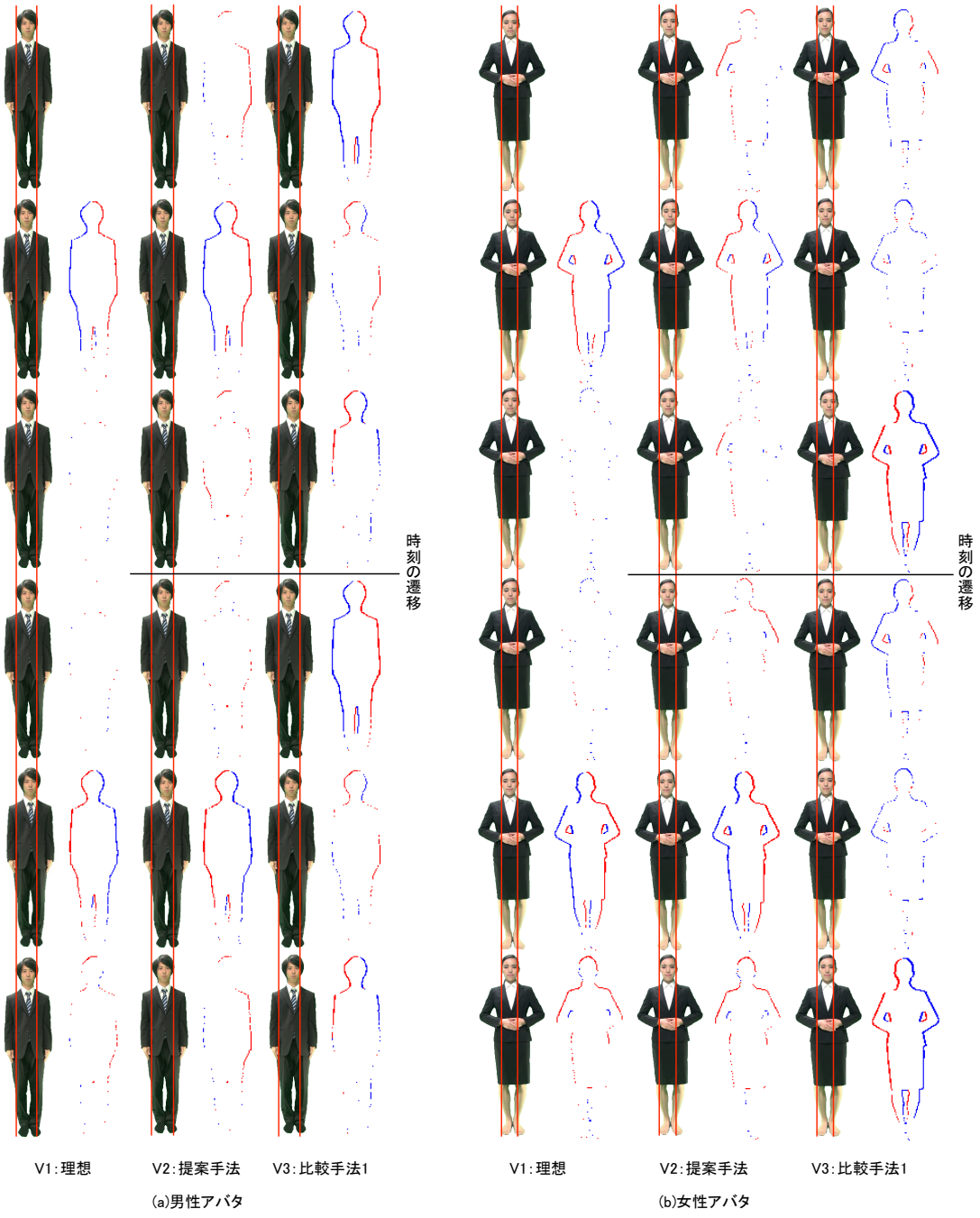


図 11　主観評価に用いた映像と振動量の可視化の例．赤色は人物から背景に変化した領域を表し青色はその逆を表す．
Fig. 11　Examples of video sequences and visualizations of movements.

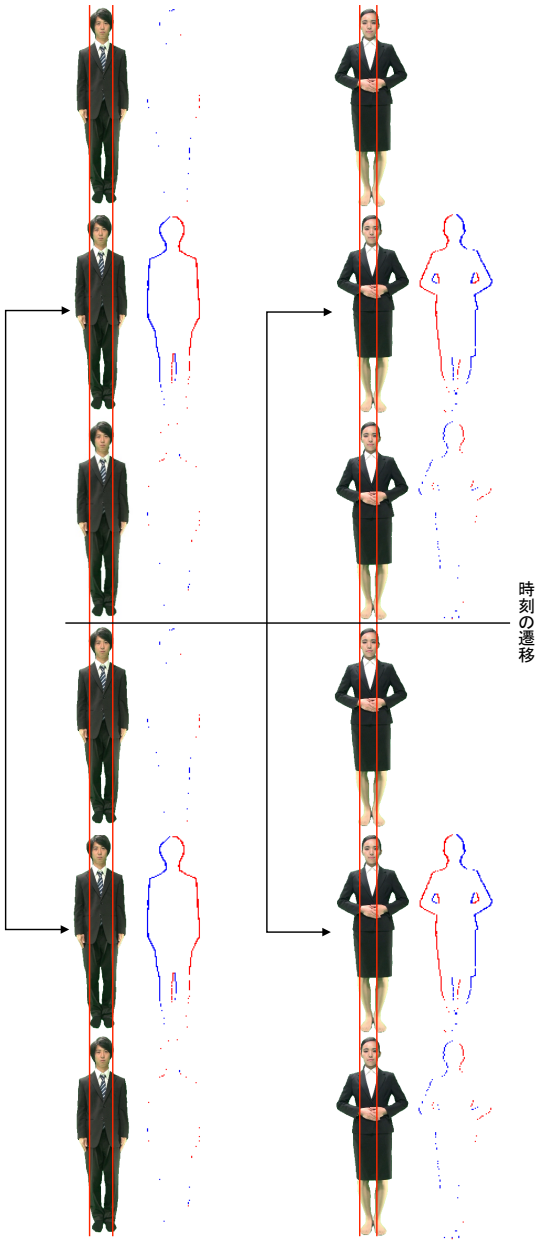


図 12　単純遷移で生成した映像と振動量の可視化の例.
Fig. 12　Examples of video sequences and movements of simple transitions.

たという意見がでた．なお，V2 と V3 を比較した際の得票数は V2 が 17 票で V3 が 5 票であった．一方，V2 は遷移時に振動量が残り，V1 と比べると不連続が発生していた．これは図 11 の上から 3 番目と 4 番目の振動量に注目すると，V1 と比べて V2 の方が僅かに多いことから分かる．このため V1 の主観評価の結果が V2 より高かったと考えられる．以上より，身体動揺を提案手法で再現することで，実写アバタの動きが人間に近づくことを確認した．

4. 4. 2　ランダム遷移と単純遷移の比較

ランダム遷移を用いて身体動揺を生成する有効性について評価した．ここでは 2 つの参照時刻で挟まれる短区間を繰り返す単純遷移と比較した．4. 4. 1 で述べたように，10 秒間のアバタ映像には男女のアバタともに 4 個の参照時刻が含まれており，参照時刻で挟まれる短区間は 3 個存在した．それぞれの短区間の時間長は，男性アバタが 3.1 秒，2.0 秒，2.9 秒，女性アバタが 4.0 秒，1.6 秒，1.7 秒であった．図 13 に各短区間における頭部の振動量 d_i の時間変化を示す．最も時間が長い短区間を繰り返す場合を単純遷移 1，その短区間と次に時間が長くかつ揺れの方向が異なる短区間を交互に繰り返す場合を単純遷移 2 とした．なお，女性アバタの 4.0 秒の短区間には瞬きの影響で左右の揺れが含まれていたため，単純遷移 1 および単純遷移 2 には用いなかった．それ以外の短区間は左右どちらかの振れが含まれていた．図 12 に単純遷移 1 で生成したアバタ映像の例と振動量の可視化の例を示す．この図より，遷移時の振動量が小さく映像は滑らかに繋がっているが，同じ動きが繰り返されていることが分かる．

ランダム遷移と単純遷移 1 と単純遷移 2 で生成した 30 秒間の映像を 11 名の被験者に見せ，人間的な動きに近い方を選択させた．図 14 にランダム遷移と比較した時の単純遷移 1 と単純遷移 2 の得票数を示す．その結果，ランダム遷移の得票数が，単純遷移 1 と比べて大幅に多かった．単純遷移 2 と比べるとランダム遷移が僅かではあるが得票数は多かった．以上のことから同じ揺れを単純に繰り返し再生するより，複数の参照時間をランダムで遷移させる方が人間的な動きに近づくことを確認した．

4. 4. 3　各類似度の有効性の検証

参照時刻を抽出するために用いた各類似度の有効性を評価した．統合類似度 s_t の場合，人物領域の形状の類似度 $s_{t,s}$ のみの場合，人物領域の見え方の類似度 $s_{t,a}$ のみの場合，顔の見え方の類似度 $s_{t,f}$ のみの場合で生成した映像についてサーストンの一対比較を用いて主観評価を行った．全ての場合において 4. 4. 1 で用いた 10 秒間のアバタ映像から 30 秒間の身体動揺の映像を生成した．統合類似度を算出する際の

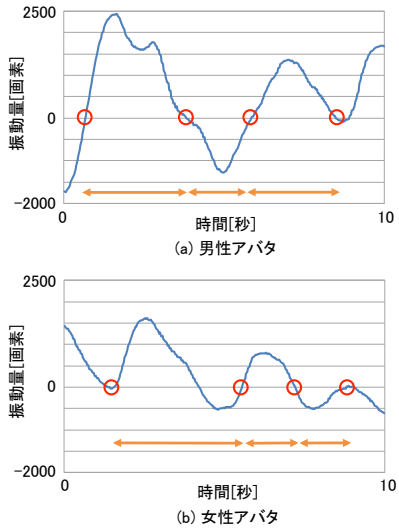


図 13　各短区間における振動量の時間変化．丸印は参照時刻，矢印は短区間を表す．
Fig. 13　Time change of movement for avatar video sequences.

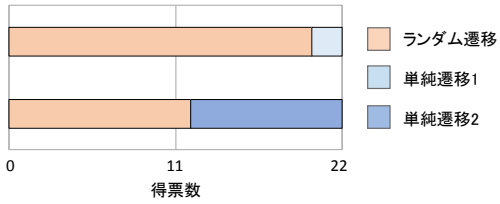


図 14　ランダム遷移と単純遷移の主観評価の得票数．
Fig. 14　Comparison of the number of votes between our method and the alternative methods.

α, β, γ は 4.4.1 と同じとした．生成前の 10 秒間の映像において各類似度の平均値と標準偏差を算出したところ s_t は 0.967 ± 0.010，$s_{t,s}$ は 0.987 ± 0.008，$s_{t,a}$ は 0.970 ± 0.008，$s_{t,f}$ は 0.912 ± 0.085 であった．

主観評価の結果を図 15 に示す．この結果より，$s_{t,a}$ を用いた場合は，$s_{t,s}$ や $s_{t,f}$ を用いた場合と比べてスコアが高くなることが分かった．また，s_t を用いた場合のスコアは，各類似度を単体で用いた場合と比べて高くなることが分かった．主観評価のスコアと α, β, γ の大小関係は同じであった．人物領域の見え方を重視し，かつ，類似度が取り得る値のスケーリング効果を考慮することが α, β, γ の設定に重要であると考えられる．以上より，各類似度を統合して参照時刻を選択

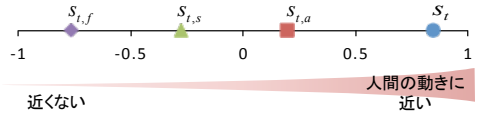


図 15　各類似度を用いた場合の主観評価の結果．
Fig. 15　Subjective assessment using each similarity.

することの有効性を確認した．

5. ま と め

本論文では，待ち状態の実写アバタで人間に近い動きを再現するために，身体動揺を計測し再現する手法について述べた．体の部位毎に振動量の時間変化を計測することで，身体動揺の中心となる参照時刻は映像中で複数回出現し，参照時刻の間には身体の揺れが存在することを確認した．これらの特徴を利用し参照時刻をランダムに遷移することで，任意の時間長の身体動揺を再現する手法について述べた．主観評価により，提案手法は比較手法と比べて人間の身体動揺に近いことを確認した．今後の課題として，待ち状態から行動状態へ滑らかに遷移する手法の開発や実稼働するインタラクションシステムでの評価，身体動揺を計測した被写体とは別人の実写アバタでその揺れを再現する手法の検討，顔全体ではなく参考文献 [23] のように表情の基本単位に注目した類似度の設計などが挙げられる．

文　献

[1] 大浦圭一郎, 山本大輔, 内匠逸, 李晃伸, 徳田恵一．キャンパスの公共空間におけるユーザ参加型双方向音声案内デジタルサイネージシステム．人工知能学会誌, Vol. 28, No. 1, pp. 60–67, 2013.

[2] 高橋範子, 小野光貴, 渡辺富夫, 石井裕．看護実習生-患者役アバタを介した看護コミュニケーション教育システム．人間工学, Vol. 50, No. 2, pp. 84–91, 2014.

[3] S. Robinson, D. Traum, I. Ittycheriah, and J. Henderer. What would you ask a conversational agent? observations of human-agent dialogues in a museum setting. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation*, pp. 28–30, 2008.

[4] R. Artstein, D. Traum, O. Alexander, A. Leuski, A. Jones, K. Georgila, P. Debevec, W. Swartout, H. Maio, and S. Smith. Time-offset interaction with a holocaust survivor. In *Proceedings of the 19th International Conference on Intelligent User Interfaces*, pp. 163–168, 2014.

[5] A. Jones, J. Unger, K. Nagano, J. Busch, X. Yu, H. I. Peng, O. Alexander, M. Bolas, and P. Debevec. An automultiscopic projector array for interactive digital humans. In *ACM SIGGRAPH 2015 Emerging*

Technologies, p. 6:1, 2015.

[6] 原健太, 堀磨伊也, 武村紀子, 岩井儀雄, 佐藤宏介. 実画像アバタを用いた対人インタラクションシステムの構築. 電気学会論文誌 C, Vol. 134, No. 4, pp. 102–111, 2013.

[7] 産業技術総合研究所人間福祉医工学研究部門. 人間計測ハンドブック. 朝倉書店, 2003.

[8] J. Gratch, J. Rickel, E. Andre, J. Cassell, E. Petajan, and N. Badler. Creating interactive virtual humans: some assembly required. *IEEE Intelligent Systems*, Vol. 17, No. 4, pp. 54–63, 2002.

[9] W. Mattheyses and W. Verhelst. Audiovisual speech synthesis: An overview of the state-of-the-art. *Speech Communication*, Vol. 66, pp. 182–217, 2015.

[10] R. Anderson, B. Stenger, V. Wan, and R. Cipolla. Expressive visual text-to-speech using active appearance models. In *Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3382–3389, 2013.

[11] D. Okwechime, E. Ong, A. Gilbert, and Richard Bowden. Social interactive human video synthesis. In *Proceedings of the 10th Asian Conference on Computer Vision*, pp. 256–270, 2010.

[12] Q. Shi, S. Nobuhara, and T. Matsuyama. Augmented motion history volume for spatiotemporal editing of 3-d video in multiparty interaction scenes. *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 25, No. 1, pp. 63–76, 2015.

[13] P. Huang, M. Tejera, J. Collomosse, and A. Hilton. Hybrid skeletal-surface motion graphs for character animation from 4d performance capture. *ACM Transactions on Graphics*, Vol. 34, No. 2, pp. 17:1–17:14, 2015.

[14] A. Egges, T. Molet, and N. Magnenat-Thalmann. Personalised real-time idle motion synthesis. *Proceedings of 12th Pacific Conference on Computer Graphics and Applications*, pp. 121–130, 2004.

[15] 山本昌彦, 吉田友英. 重心動揺計を用いた体平衡機能検査： 重心動揺検査・電気性身体動揺検査. *Equilibrium research*, Vol. 70, No. 3, pp. 135–144, 2011.

[16] 竹内弥彦, 下村義弘, 岩永光一, 勝浦哲夫. 小型三軸加速度計による高齢者の動的バランス評価の有用性. 理学療法科学, Vol. 22, No. 4, pp. 461–465, 2007.

[17] 大西智也, 橘浩久, 武田功. 安静な立位における足位の違いが下肢帯および下肢の動揺に及ぼす影響. 理学療法科学, Vol. 30, No. 2, pp. 313–316, 2015.

[18] F. Wang, M. Skubic, C. Abbott, and J. M. Keller. Body sway measurement for fall risk assessment using inexpensive webcams. In *Proceedings of International Conference of Engineering in Medicine and Biology Society*, pp. 2225–2229, 2010.

[19] L.F. Yeung, Kenneth C. Cheng, C.H. Fong, Winson C.C. Lee, and Kai-Yu Tong. Evaluation of the microsoft kinect as a clinical assessment tool of body sway. *Gait & Posture*, Vol. 40, No. 4, pp. 532–538, 2014.

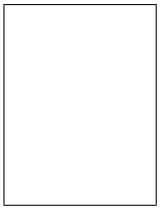
[20] J. M. Saragih, S. Lucey, and J. F. Cohn. Deformable model fitting by regularized landmark mean-shift. *IEEE international Journal of Computer Vision*, Vol. 91, No. 2, pp. 200–215, 2011.

[21] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, Vol. 13, No. 4, pp. 600–612, 2004.

[22] L. L. Thurstone. A law of comparative judgment. *Psychological Review*, Vol. 34, No. 4, pp. 273–286, 1927.

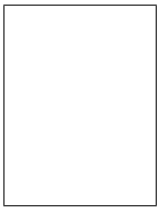
[23] 竜己坂口, 寛山田, 繁生森島. 顔画像を基にした 3 次元感情モデルの構築とその評価. 電子情報通信学会論文誌. A, 基礎・境界, Vol. 80, No. 8, pp. 1279–1284, 1997.

（平成 xx 年 xx 月 xx 日受付）



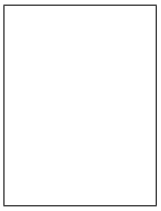
宮内　翼

2016 年 鳥取大学大学院工学研究科情報エレクトロニクス専攻在学中.



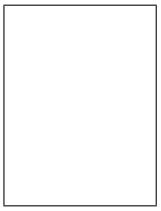
吉村　宏紀　（正員）

1993 年鳥取大学工学部知能情報工学科卒業．1997 年鳥取大学大学院工学研究科博士後期課程情報生産工学専攻修了．博士 (工学)．現在鳥取大学大学院工学研究科情報エレクトロニクス専攻，助教．信号処理の研究に従事．情報処理学会の会員．



西山　正志　（正員：シニア会員）

2002 岡山大学大学院博士前期課程了．株式会社東芝研究開発センターを経て，現在鳥取大学大学院工学研究科准教授．学際情報学博士（東京大学）．画像認識，拡張現実感の研究に従事．2009 山下記念研究賞などを受賞．情報処理学会会員



岩井　儀雄　（正員）

1997 年 (平成 9 年) 大阪大学大学院基礎工学研究科博士課程後期修了．同年同大学院助手．2003 年 (平成 15 年) 同大学院助教授．2004 年 (平成 16 年)5 月〜2005 年 (平成 17 年)3 月英国ケンブリッジ大学客員研究員．2007 年 (平成 19 年) 同大学院准教授．2011 年 (平成 23 年) 鳥取大学大学院工学研究科教授．パターン認識の研究に従事．博士（工学）