Vol.2013-CVIM-187 No.42
2013/5/30

顔画像の適応型疎テンプレート追跡に基づく母音認識用固有空間の作成

瀬戸 寛之[1]　福永 嵐馬[1]　尺長 健[1]

概要：本稿では，顔画像の適応型テンプレート追跡に基づくオンライン学習によって生成された母音認識用固有空間の作成法を提案するとともに，従来，顔追跡・認識融合系による人物識別で用いられてきた固有空間と加重方程式を用いた識別法を，母音認識に適用することで，自然環境下でのリップリーディングを実現する方法を検討する．また，実データによるオンライン学習で得られた固有空間を用いて顔追跡・認識融合系の中に母音認識系を構成した結果について報告する．

1. はじめに

顔画像には，様々な情報が含まれており，これを有効に利用することは今後の人間と計算機の間の自然なユーザインタフェースを実現するためにも重要な課題である．我々の研究室では，疎テンプレートマッチングに基づく顔画像追跡・認識の研究を幅広く進めてきたが，これらの研究は大きく２種類に分類され，それぞれ次のような研究経過をたどってきた．

まず，第１番目は追跡の安定化を目指したものであり，松原・尺長 [1] では単純な疎テンプレート追跡であったものが，テンプレートの変化に対応するために，固有テンプレート [2] と適応型テンプレート [3] に分化した．これらの研究は，坂部・田口・尺長 [4]，瀬戸・田口・尺長 [5] によって，適応型テンプレート追跡を利用した固有テンプレートのオンライン学習へと展開してきた．

一方，２次元テンプレートから３次元テンプレート追跡への展開は，まず，岡・黒田・右田・尺長 [6] によって行われ，粗密探索による６次元姿勢空間の効率的探索により，追跡系が安定化したのを契機として，追跡系と認識系の融合を目指した研究が進んだ．岡・尺長 [7], [8] は，正規化画像によって構成される固有空間を形状推定に利用する方法を提案し，照明変動・姿勢変動が存在する動画中の人物認識と形状推定を実時間で実現することに成功している．この系では，人物認識と形状推定を加重方程式と呼ぶ線形連立方程式に帰着し，２５人規模の系で実時間認識が可能であることを示している．最近，この研究の改良版として，中岸・尺長 [9], 中岸・岡・尺長 [10] は，それぞれ，階層的加重方程式，および，加重方程式の並列不足決定系を構成することで，より高速に安定した人物認識系を１００人規模で実現できることを示している．これらの追跡・認識融合系は人物認識ばかりでなく，表情認識などへの展開が可能であると考えられる．

顔画像情報処理の研究は，今後，ますます高度化していくと考えられる．本稿では，顔を取り巻く情報処理の重要な例題として，リップリーディングを取り上げ，オンライン学習と顔追跡・認識系を有効に組み合わせることを検討する．例題として，日本語発話時の母音認識を，顔追跡・認識融合系を用いて実現する問題を取り上げる．ここで，日本語発話中の各母音発声時の画像パターンを効率的に学習するため，適応型テンプレート追跡を利用したオンライン学習を利用する．また，この方法によって効率よく構成された固有空間を顔追跡・認識融合系で用いることにより，自然環境下で動作する母音認識系を構成することを目指す．

第 2 節では, 適応型テンプレート追跡を利用した固有テンプレートのオンライン学習の概要を述べるとともに，母音固有空間構成に特有の問題を分析する．第 3 節では，追跡・融合系の概要と母音認識問題への適用法を示す．また，母音認識における加重方程式の取扱いについてのべる．第 4 節では，これらの議論を実動画に適用した実験結果を述べる. 即ち，オンライン学習によって構成した母音固有空間を用いることによる母音認識性能を検証するとともに，今後の課題をまとめる．

[1] 岡山大学
Okayama University

2. 適応型追跡に基づく見え変化の学習

2.1 3 次元 WSL モデル

野口・尺長 [3] は, WSL モデル [11] と疎テンプレート追跡 [1] を組み合わせた適応型疎テンプレート追跡を提案している. WSL モデルは, ある時刻 $t-1$ にデータ d_{t-1} をとったとき, 次時刻 t においてデータ d_t をとる確率 $pt(d_t|d_{t-1})$ を混合ガウス分布で表現したものである. WSL は 3 つの要素 (Stable, Wandering, Lost) で構成され，各要素のガウス分布は $p_s(d_t|\mu_{s,t}, \sigma^2_{s,t})$, $p_w(d_t|d_{t-1})$, $p_l(d_t)$ で表される. これらの要素からなる d_t の確率密度を次式で表す.

$$p(d_t|\mathbf{q}_t, \mathbf{m}_t, d_{t-1}) = m_{w,t}p_w(d_t|d_{t-1}) + m_{s,t}p_s(d_t|\mathbf{q}_t) + m_{l,t}p_l(d_t) \quad (1)$$

ただし，$\mathbf{q}_t = (\mu_{s,t}, \sigma^2_{s,t})$，$\mathbf{m}_t = (m_{w,t}, m_{s,t}, m_{l,t})$ であり，$\mathbf{m}_t$ は各要素に対する重みである. このモデルパラメータ $\mathbf{q}_t$，$\mathbf{m}_t$ の更新と姿勢推定の繰り返しにより,Jepson ら [11] は対象の見えの逐次学習と追跡を実現しているが, この方法では姿勢推定の計算コストが大きく, 実時間追跡は向いていないことが分かった. そこで野口・尺長 [3] は, WSL モデルにおける学習に疎テンプレート追跡を統合し, 疎テンプレート追跡による高速な姿勢推定と, 疎テンプレート追跡によって得られた切り出し画像を d_t としてモデルパラメータを更新することで, 高速な逐次学習追跡を実現している.

一方, 岡ら [6] は, 顔の画像情報と 3 次元形状を組み合わせた 3 次元テンプレートに, 正規化固有空間を組み合わせることで, 3 次元疎テンプレート追跡を実現している. この 3 次元疎テンプレート追跡において, 各表面パッチを WSL モデルで形成することで, 3 次元適応型疎テンプレート追跡を実現する. なお, 本稿では WSL モデルを顔の画像情報にのみ適用し, 形状は固定として考える. 形状に対する WSL モデルの適用は今後の課題である. 3 次元 WSL モデルの更新例を図 1 に示す. 図中, 左列は入力画像を示し, 右側に固定形状表面のテクスチャの変化 (WSL モデルの S 値) の様子を示す.

2.2 3 次元 WSL モデルによる適応型追跡

適応型疎テンプレート追跡 [3] は前述のとおり, WSL モデルの姿勢推定を疎テンプレート追跡によって行い, 疎テンプレート追跡によって得られた推定姿勢の画像を基にモデルパラメータを更新する方法である. ここで, WSL モデルの更新によって作成される密なテンプレートを基に, 次フレームで使用する疎テンプレートを作成する. WSL モデルの更新によって得られる密なテンプレートは前時刻と現在の時刻との変化を捉えたものであり, WSL による逐次学習を疎テンプレートに反映することが可能となる. 姿勢推定は疎テンプレートマッチングとパーティクルフィルタからなり, 2 次元テンプレートで追跡を行う場合, 姿勢推定に必要なパラメータは 4 つ (並進:2 自由度, 回転:1 自由度, スケール:1 自由度) である.

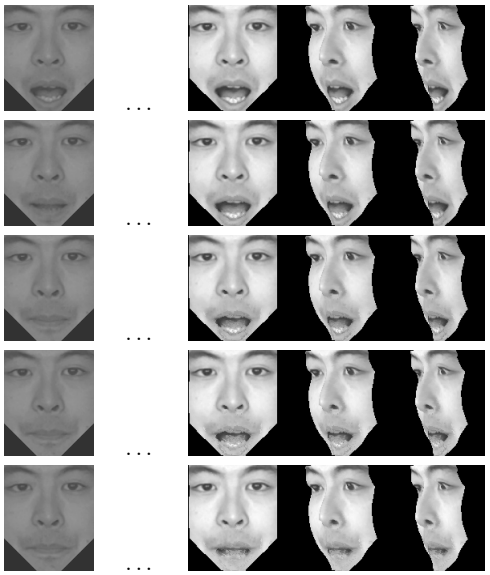


図 1　3 次元 WSL モデルの更新

この適応型疎テンプレート追跡を 3 次元に拡張することにより, 姿勢空間は 6 次元になる. 即ち, 剛体姿勢の持つ 6 自由度 (並進:3 自由度, 回転:3 自由度) をパーティクルフィルタで追跡することになる. 3 次元テンプレートによる追跡 [8] も 2 次元の場合と同様, 疎テンプレート追跡とパーティクルフィルタから成る. 本稿では 3 次元テンプレートの形状は固定で, テクスチャのみを WSL モデルによって更新する.

2.3 各母音に対応する固有空間のオンライン学習

坂部ら [4] により, 適応型疎テンプレート追跡による見え変動の学習が提案されており, 疎固有テンプレート追跡に有効な固有空間の学習を実現している. また, 瀬戸ら [5] は, 動画内で動き回る人物の見え変動を方向別に学習することで, 見えごとの情報を学習する方法を提案している. 本稿の母音学習では, 適応型疎テンプレート追跡を用いて母音ごとに見え変化を学習・固有空間を作成する方法を検討する.

本稿では追跡を簡略化するため, 初期姿勢・初期テンプレートを与えて追跡を開始する. 初期テンプレートに使用する初期画像を $\mathbf{x_d}$ とする. まず, 学習したい母音 v の画像 $\mathbf{X}_{v0}$ を得るまで適応型疎テンプレート追跡を行い, 推定姿勢から得られた切り出し画像 $\mathbf{X}_t$ の正規化画像 $\mathbf{x}_t$ と $\mathbf{x}_d$ の正規化相関を計算する. 相関が閾値を越えたとき, $\mathbf{X}_t$ を学習画像集合 $\{\mathbf{X}_v\}$ に加える.

新たに学習画像を得た時点で, 学習画像集合 $\mathbf{X}_v$ を基に

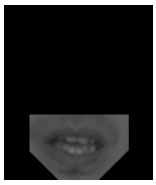


図 2　学習に使用する領域 $\mathbf{Q}$

正規化固有空間 $(\Phi\ \bar{\mathbf{x}}_v)$ を作成し, 以降の学習に使用する. 本稿の学習は発音に伴う口の見えの変化を重視し, 画像の使用領域 $\mathbf{Q}$ を口元に制限する. 図 2 に領域 $\mathbf{Q}$ をしめす. 時刻 t の姿勢推定から得られた正規化画像を $\mathbf{x}_t$, 同次固有空間を $\tilde{\Phi} = [\Phi\ \bar{\mathbf{x}}_v]$ で表す. 得られた正規化画像の部分画像 $\mathbf{Q}\mathbf{x}_t$ の同次固有空間 $\tilde{\Phi}$ への部分射影は $\mathbf{x}_t^* = (\mathbf{Q}\tilde{\Phi})^+\mathbf{x}_t$ であり, 部分画像 $\mathbf{Q}\mathbf{x}_t$ と射影から得られる同じ部分の画像 $\mathbf{Q}\tilde{\Phi}\mathbf{x}_t^*$ との正規化相関 $C(\mathbf{Q}\mathbf{x}_t, \mathbf{Q}\tilde{\Phi}\mathbf{x}_t^*)$ を求める. 相関が閾値を越えた場合, 新たな見え変化が発生したと判定し, $\mathbf{x}_t$ を学習画像集合に加え, 固有空間を更新する.

3. 母音認識への応用

3.1 追跡・認識融合系

本稿の追跡・認識は岡・尺長 [7], [8] の追跡・認識融合系にならう. 岡・尺長 [7], [8] は, 3 次元疎固有テンプレート追跡 [6] と, 後述する加重方程式による人物認識からなる. 3 次元疎固有テンプレート追跡は, 固定形状を対象とした 3 次元疎テンプレートマッチングを用いて, 6 次元姿勢空間においてパーティクルフィルタによる姿勢追跡を行う方法である. これにより, 各フレームにおいて 3 次元モデルが更新可能となる. なお, 文献 [7], [8] では顔の形状推定を同時に行っているが, 本稿では前述のとおり, テクスチャの更新のみ行う.

3.2 母音固有空間と個別母音固有空間

岡・尺長 [7], [8] は, 追跡結果から認識を行う際, 固有空間に各人物の個人固有空間を構成し, 加重方程式を過剰決定系で解く方法を提案している. 本稿の母音認識はこれを基にし, 学習によって得られた母音別発音画像から 2 種類の固有空間を作成する.

学習によって得られた各母音画像集合 $\mathbf{X}_v$ を統合した画像集合を $\mathbf{x}$ とする. この $\mathbf{x}$ を基に作成する正規化固有空間 $(\Phi, \bar{\mathbf{x}})$ を母音固有空間と呼ぶ. このとき, 各画像集合 $\mathbf{x}_v$ 内の母音画像 $\mathbf{x}_{vi}$ の同次固有空間への射影は次式で表される. ここで, $i(i = 1, \ldots, I)$ は学習画像集合に含まれる画像の番号である.

$$\tilde{\mathbf{s}}_{vi} = (\mathbf{P}\tilde{\Phi})^+(\mathbf{P}\mathbf{X}_{vi}) \tag{2}$$

$$\hat{\mathbf{s}}_{vi} = [\alpha\mathbf{s}_{vi}^T\ \alpha]^T \tag{3}$$

これにより, 射影の正規化表現 $\hat{\mathbf{s}}_{vi} = [\alpha\mathbf{s}_{vi}^T, \alpha]^T$ を獲得する. ここで各母音 v について, $\mathbf{S}_v = [\mathbf{s}_{vi} | i = 1, \ldots, I]$ を主成分分析することで, 固有空間 $< \bar{\mathbf{s}}_v, \nu_v >$ を得る. この空間を個別母音固有空間と呼ぶ. 任意の画像 $\mathbf{X}$ の正規化固有空間への射影 $\mathbf{s}$ が与えられたとき, $\mathbf{s}$ の個別母音固有空間 $< \bar{\mathbf{s}}, \nu_v >$ への射影は次式となる.

$$\mathbf{s}_v = \nu_v\nu_v^T(\mathbf{s} - \bar{\mathbf{s}}_v) + \bar{\mathbf{s}}_v \tag{4}$$

岡・尺長 [7], [8] では, 個人固有空間への射影が同様の式で表されている. 即ち, 本稿での個別母音固有空間への射影は, 文献 [7], [8] と形式的には同じ式で実現される. 一方, 文献 [7], [8] で式 (4) は照明適応 (photometric adjustment) と呼ばれているが, 本稿の式 (4) は照明適応ではない.

3.3 加重方程式による母音認識

母音の認識には加重方程式 [9] を利用する. ある画像 $\mathbf{X}$ が与えられたとき, 最初に母音固有空間への射影 $\mathbf{s}$ が求められる. ここから, $\mathbf{s}$ を個別母音固有空間へと射影し, $\mathbf{s}_v$ を得る. $\mathbf{s}_v$ は即ち, 入力画像に対応する各母音の代表の 1 点を選択していることになる. 各母音の代表の 1 点が定まることから, 加重方程式では各母音の 1 つの画像 $\mathbf{s}_v$ のみ考えて加重方程式を解くことになる. つまり, $\mathbf{s}_v$ の加重方程式により, 入力画像による人物の発音認識が実現される. ここで, 加重方程式は次式で表される式である.

$$\hat{\mathbf{S}}_V\mathbf{w} = \hat{\mathbf{s}} \tag{5}$$

この $\mathbf{w} = [w_1, \ldots, w_5]^T$ が求めるべき加重ベクトルであり, $\hat{\mathbf{S}}_V$ は, 5 つの母音の $\mathbf{s}_v$ の拡張表現を並べたものである.

$$\hat{\mathbf{S}}_V = [\hat{\mathbf{s}}_1 \ldots \hat{\mathbf{s}}_5] = \begin{bmatrix} \mathbf{s}_1 & \ldots & \mathbf{s}_5 \\ 1 & \ldots & 1 \end{bmatrix} \tag{6}$$

過剰決定系では, 式 (6) の最小二乗誤差を与える $\mathbf{w}$ が得られる. また, ここで $\mathbf{w}$ は $\mathbf{1}^T\mathbf{w} = 1$ という制約のもとで最適化されるため, 加重方程式の解は $\mathbf{s}$ を $\mathbf{s}_v$ の重み付き平均で表すこととなる. 即ち, 加重が最も大きい発音とすることができるため, 発音認識は次式となる.

$$V_{max} = \arg\max_v\ w_v \tag{7}$$

4. 実験

4.1 実験条件

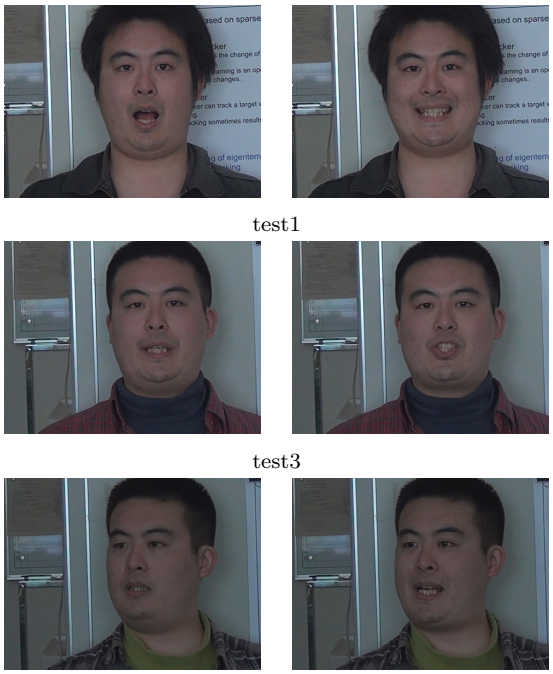
実験に使用する動画像の条件を表 1 に示す. 表中のシーケンスの項目にはシーケンス名とフレーム数を記載しており, カッコ内の数字はシーケンス中で発話が行われているフレーム数を表す. 基本的に発話をしている人物を撮影したものであるが, 内容や状態の違いを持たせている. まず, test1, 2 は人物が正面を向いたまま「あ・い・う・え・お」と繰り返し発音するシーケンス, test3, 4, 5 は人物が正面を向いたまま自然な発話を行うシーケンス

表 1　使用動画の撮影条件

撮影器具	SONY HANDYCAM HDR-TD10
解像度	1920x1080
フレームレート	60fps

表 2　シーケンスデータ

シーケンス名	フレーム数 (母音フレーム数)
test1	1780(897)
test2	2394(1070)
test3	730(332)
test4	677(376)
test5	1584(732)
test6	686(242)



test1

test3

test6

図 3　使用画像の一部

である. test6 は発話に加えて頭部の動きを加えたもので, test6 は一定方向を向いたまま発話するシーケンスである.test1, test3, test6 の例を図 3 に示す. 使用する各シーケンスには予め音韻ラベルが与えられており, 各フレームでどの母音を発音しているかが分かるようにした. ここで, 子音については本稿では取り扱わず, また, 母音間であっても音韻が変化する過程 (「あ」から「い」など) では正確なラベルが与えられないため, 「該当なし」を示すラベルを設定している.

4.2　学習実験

まず, WSL モデルの α を変更させた場合, 学習結果に変化が発生するかを調べた. ここで α は学習率と呼ばれ, WSL モデルの更新において, 前のテンプレートからの変化をどれだけ反映させてモデルを更新するかを表すパラメータである. この α を 0.0 から 0.5 まで変化させ, 3 次元適応型疎テンプレート追跡による学習に影響が生じるか, またどのような影響が出るかを検証した. なお, 閾値は 0.999 とした. また, 人物の発音に伴う口や目元の変化により推定姿勢にズレが生じる可能性があるため, 姿勢推定に用いる領域を図 4 で黒塗りにしていない部分に制限して実験を行った.

図 4　適応型追跡における領域制限

学習結果を表 3 に, 学習画像の一部を図 5 に示す. 表 3 によると, α を変更すると多少の変動はあるものの, ほぼ一定した枚数を学習していることがわかる. また, 図 5 に示すように, 品質の良い画像を学習できていることが確認された. しかし, 中には位置ズレを含む画像も含まれており (図 6), 適応型追跡には精度改善の余地があることがわかる.

次に, 学習の閾値変化による影響を調べた. 閾値を変化させることでより多くの情報を持った固有空間を作成できると考えられ, 学習の閾値を 0.9990 から 0.9999 まで変化させて, 学習した画像から固有空間を作成し, どのような差異が生じるかを検証した. なお, 学習率は 0.3 とした. 表 4 に学習した画像の枚数を示す. 表から分かるように, 閾値を高く設定することで, より多くの母音画像を学習できる. 閾値 0.9990 で学習した枚数と 0.9999 で学習した枚数を比較すると, どの学習パターンでも 3, 4 倍の画像を学習している. ここで, 学習の閾値を高めることにより, 微細な変動を学習することが期待でき, 連続した変化を伴う発音の認識に必要なデータをより多く獲得できると考えられる. しかし, 学習した画像すべてに必ずしも有効な情報が入っているとは限らない. 閾値を 0.9999 に設定した場合, 学習する画像の枚数は膨大であるが, その中には位置ズレを伴って学習された画像も含まれていると考えられる.

各画像集合から作成した累積寄与率 0.95 の母音固有空間の次元数の変化を表 4 のカッコ内に示す. test1, test2 を見ると, 閾値上昇によって学習画像が増えるため, 固有空間の次元数は上がるが, 累積寄与率が 0.95 となる次元数の変化はほとんどないことがわかる. また, これらのシーケンスでは閾値の上昇による累積寄与率 0.95 水準の次元数の変化は少ないと言える. test3, test4 を見ると, 閾値の上昇に伴って累積寄与率 0.95 水準の次元数も上昇している. これは, 発音に伴う口元の微小な変化を多く捉

表 3　学習率変更による学習枚数の比較

学習率	0.0	0.1	0.2	0.3	0.4	0.5
test1	81	95	83	81	82	85
test2	84	109	110	113	109	108
test3	95	103	95	94	94	92
test4	61	73	73	78	76	75

表 4　閾値による学習枚数と母音固有空間の次元数 (カッコ内の数は累積寄与率 0.95 となる次元数を示す)

閾値	0.9990	0.9992	0.9994	0.9996	0.9999
test1	81 (29)	95 (31)	114 (32)	153 (33)	341 (32)
test2	113 (39)	130 (41)	157 (43)	197 (43)	376 (34)
test3	94 (42)	112 (45)	136 (49)	177 (52)	305 (54)
test4	78 (40)	88 (42)	119 (48)	160 (54)	338 (63)

図 5　学習画像の一部 (上から「あ」,「い」,「う」,「え」,「お」の順)

えているということであり, 発話シーケンスから多くの情報を得ていると考えられる. しかし, 適応型追跡の姿勢推定精度を考えると, 位置ズレによる変化など認識には不要な情報を学習している可能性もある. そのため, 安易に閾値を高く設定するのはふさわしくないと考えられる. これらの結果を総合的に判断し, 以下の実験では WSL モデルの学習率 α を 0.3, 閾値を 0.999 に設定する.

4.3　認識比較実験

学習実験から得られた個別母音固有空間と, 各個別母音固有空間の学習に用いられた画像すべてから構成される固有空間 (母音固有空間) を用いて, テスト動画像上で認識実験を行った. 具体的には, 人物の発話シーケンスの追跡・認識を行い, 発音が生じているフレームでその発音がどの母音であるかを認識させる実験を行った. 実験では, test3・test4 から学習した画像を使用し, test3・test4・test5・test6 に疎固有テンプレート追跡・母音認識を行い, 正確な母音認識が行えるかを検証した. 認識に使用する個別母音固有空間の一部を図 7 に示す. 実験では, test3・test4 から得られた学習画像 163 枚から, 累積寄与率 0.95 となる母音固有空間を作成して認識を行った. なお, 学習元となった test3, test4 には固有空間の学習に用いたフレームが存在するため, 認識対象から除外した上で実験を行った. 母音認識は音韻ラベルが振られているフレームのみで行い, 認識率は正しく認識した発音数/全体の発音数によって算出した. また, 実際の発話は連続的であり口の形状が変化すること, 前後の子音によって母音の発音が影響を受けることを考慮し, 第 2 位累積分類率を合わせて求めた. 使用した疎固有テンプレート追跡の各種パラメータを表 6 に示す.

表 5 に認識結果を示す. まず学習元となった test3, test4 では認識率は約 80%, 第 2 位累積分類率は約 95%前後となった. 学習したフレームを除外しての認識実験だが, これらの画像は固有空間に必要な要素として学習されなかった画像, 即ち固有空間で十分対応可能な画像であるため, 学習元シーケンスの認識率は高くなったと考えられる. 続いて学習に使用していない test5, test6 に認識を行うと, 認識率・第 2 位累積分類率ともに低い数値を記録した.test5 は test3, test4 と同じく, 正面を向いたまま発話しているシーケンスではあるが, 同様に正面を向いて発話している test3, test4 から得た学習画像で認識しているにも関わらず認識率が低くなっている. 表 7, 8 に認識結果の内訳を示す. 表を見ると,「え」を「い」と認識,「お」を「う」と認識しているケースが多い. 事前に音韻ラベルをつける際, 発生している母音に合わせてラベルを与えたため, 自然な発話を行うシーケンスである test3, test4 から学習した固有空間に, 本来別の発音と認識すると思われる画像が含まれていたため, 誤認識を引き起こしたと考えられる. test6 は作成した母音固有空間による追跡が不安定なために充分な切り出しが行えず, 誤認識を多発したと考えられる.

先ほどの実験では, 1 フレームの切り出し画像を用いて認識を行ったが, 発話とは本来連続したシーケンスであり, 1 フレームだけではなく数フレームに渡る見えの変化・音声を伴うものである. そこで, 連続する 2 フレームで母音認識を行った. 認識結果を表 9 に示す. この実験では test6 を除外した. 実験の結果, 認識率にわずかな変化が見られたが大幅な改善は見られなかった・

図 6　位置ズレ状態で学習した例

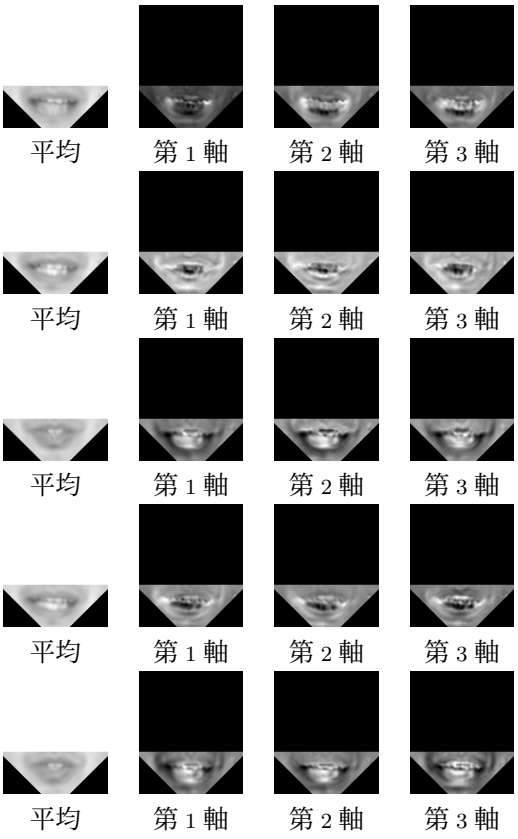


図 7　母音認識に使用する個別母音固有空間の一部. 上から「あ」,「い」,「う」,「え」,「お」の順

表 5　母音認識率・第 2 位累積分類率 (単一フレーム認識)

	test3	test4	test5	test6
認識率 (%)	81.3	79.7	45.5	31.8
第 2 位累積分類率 (%)	94.0	95.5	74.5	51.2

表 6　疎固有テンプレート追跡のパラメータ

サンプル数	1000
(上位サンプル)	(10)
σ	(5 5 10 5 5 2.5)
粗密探索段数	3

表 7　認識の内訳 (test3, 4. 縦軸は正解の母音, 横軸は認識結果)

	a	i	u	e	o
a	65	7	0	0	0
i	5	148	1	0	1
u	0	6	114	0	5
e	9	30	4	23	0
o	5	0	37	0	102

表 8　認識の内訳 (test5. 縦軸は正解の母音, 横軸は認識結果)

	a	i	u	e	o
a	80	77	1	3	0
i	23	113	3	5	1
u	10	44	40	0	38
e	15	94	8	7	1
o	17	14	45	0	93

表 9　母音認識率・第 2 位累積分類率 (2 フレーム認識)

	test3	test4	test5
認識率 (%)	82.5	84.4	46.8
第 2 位累積分類率 (%)	96.8	96.3	74.7

4.4　不足決定系による認識実験

前節までの認識実験では, 適応型追跡による学習から得た母音固有空間と個別母音固有空間から加重方程式を作成して認識を行った. 一方, 岡・尺長 [7], [8] には, 多重登録による認識法が対案として示されている. この方法を母音認識に適用すると, 個別母音固有空間を作成せず, 登録画像の射影すべてを使用することで, 加重方程式を解くことになる. 具体的には, 160 枚の登録画像から作成した母音固有空間 (55 次元) を考える. このとき, 加重方程式 $\hat{\mathbf{S}}_K\mathbf{w} = \hat{\mathbf{s}}$ において $m+1 < K$(m は母音固有空間の次元数, K は登録枚数) となる. すなわち, 加重方程式は不足決定系となり,$\mathbf{w}$ に関する解空間は $(K-m-1)$ 次元となる. 不足決定系においては, 擬似逆行列による解は $\mathbf{1}^T\mathbf{w} = 1$ と $\mathbf{s} = \mathbf{S}_K\mathbf{w}$ を満たす解空間内で $\mathbf{w}^T\mathbf{w}$ を最小とする $\mathbf{w}$ となり, 全登録画像の重み付き平均が得られる.

試みに test5 をテストデータとした場合の認識率を求めたところ, 認識率が 57.2% に上昇することが確認できた. この結果は, 母音をサブカテゴリとして取り扱うことで認識率の上昇が期待できることを示していると考えられる.

5. まとめ

本稿では, 従来 2 次元で使用されていた適応型疎テンプレート追跡による見え変化の学習を 3 次元テンプレートへと拡張し, 母音認識用固有空間の作成を適用した. 対象の見えを逐次学習しつつ高速な追跡を行う適応型疎テンプレート追跡を 3 次元テンプレートへ拡張し, 3 次元テンプレートによる適応型疎テンプレート追跡を実現した. この 3 次元適応型追跡を使用し, 人物の発話に伴う見えの変化を学習し, 認識に使用する個別母音固有空間の作成に使用した. 発話する学習対象に対して適応型追跡を行うため, 追跡に用いる点を顔の上半分に限定することにより口の変化による追跡への影響を抑え, 口元の画像を使用することで母音ごとの見えの変化の学習する方法を提案した.

得られた母音画像集合から, すべての画像を使用した母音固有空間を作成し, この空間に各母音画像を射影して母音ごとに主成分分析した個別母音固有空間を作成した. これら 2 種類の母音固有空間を使用し, 疎固有テンプレート追跡によって推定された姿勢の切り出し画像から加重方程式を生成し, 人物の母音認識を実現した.

実験ではまず, 適応型疎テンプレート追跡による母音固有空間の学習を行った. 適応型追跡の学習率・学習における閾値をそれぞれ変化させ, 学習に与える影響を調べた. その結果, 学習率による画像学習への影響が小さいこと, 閾値を高くすることで学習枚数は増えるが余分な学習を行う可能性があることが分かった. 次に, 自然な発話を行うシーケンスから個別母音固有空間を学習し, 得られた個別母音固有空間から母音固有空間を作成し, 別のシーケンスで追跡・認識実験を行った. 結果, 認識結果は 40% 程度となった. 自然な発話シーケンスでは発音と見えに大きく差があり, また, 発話は連続的なデータの集まりであるため, 1 枚の画像から正確に母音を認識することが困難であるためだと考えられる.

今後の課題として, 対象の動きへの対応・音韻ラベルの細分化・音韻から音韻への変化による見え変化への対応・発音を連続的データとした運用方法が上げられる. また, 画像だけでは母音認識が難しいと考えられるため, 音声認識と組み合わせての運用も上げられる.

参考文献

[1] 松原康晴，尺長健，“疎テンプレートマッチングとその実時間物体追跡への応用，ibitglue ” 情報処理学会論文誌:コンピュータビジョンとイメージメディア，vol.46, no.SIG9(CVIM 11)，pp.60-71，2005.

[2] T. Shakunaga, Y. Matsubara and K. Noguchi, “Appearance tracker based on sparse eigentemplate,” Proc. Int’l Conf. on Machine Vision & Applications, pp.13–17, 2005.

[3] 田口智行，野口清志，尺長健，“適応型疎テンプレート追跡,” 電子情報通信学会論文誌.VOL.J93-D,NO.8，pp.1502-1511，2010.

[4] K. Sakabe, T. Taguchi and T. Shakunaga, “Automatic eigentemplate learning for sparse template tracker,” Lecture Notes in Computer Science Proceedings of the 3rd Pacific Rim Symposium on Advances in Image and Video Technology, vol.16, pp.714–725, 2009.

[5] H. Seto, T. Taguchi and T. Shakunaga, “Directional eigentemplate learning for sparse template tracker,” Lecture Notes in Computer Science Proceedings of the 3rd Pacific Rim Symposium on Advances in Image and Video Technology, 2011.

[6] Y. Oka, T. Kuroda, T. Migita and T. Shakunaga, “Tracking 3d pose of rigid object by sparse template matching,” Proc. The 5th International Conference on Image and Graphics, ICIG2009, 2009.

[7] Y. Oka and T. Shakunaga, “Sparse eigentracker augmented by associative mapping to 3d shape,” Automatic Face and Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on 21-25 March 2011, pp.649–656, 2011.

[8] Y. Oka and T. .Shakunaga, “Real-time face tracking and recognition by sparse eigentracker with associative mapping to 3d shape,” Image and Vision Computing Volume 30 Issue 3, March, 2012, pp.147–158, 2012.

[9] H. Chugan and T. Shakunaga, “Hierarchical approach to weight equations in tracking and recognition framework,” The 10th IEEE Conference on Automatic Face and Gesture Recognition (FG2013), 2013.

[10] H. Chugan, Y. Oka and T. Shakunaga, “Parallel underdetermined approach to weight equations in tracking and recognition framework,” International Conference on Machine Vision & Applications (MVA2013), 2013.

[11] A.D. Jepson, D.J. Fleet and T.F. El-Maraghi, “Robust online appearance models for visual tracking,” IEEE Trans. Pattern Analysis and Machine Intelligence, vol.25, no.10, pp.1296–1311, 2003.