Vol.2016-CVIM-203 No.35
2016/9/6

自然会話における頭部動作検出

秋山 解[1,a)]　伍 洋[1]　Kris M. Kitani[2]　László A. Jeni[2]

概要：非言語情報はコミュニケーションにおいて重要だが，非言語情報の内うなずきや首振りといった頭部動作の自動認識は，これまで限定された条件下でのみ研究されてきた．本研究では自然会話における多種類の頭部動作を自動検出するシステムを目指す．そのための手段として，HOG-like な特徴量である HoVA 特徴量を提案し，自然会話映像データセット (FIPCO) によってその有効性を確認した．さらに，頭部動作検出において最先端のモデルである LDCRF を適用するにあたっては，従来考慮されていなかったオンラインテストへの応用を行い，オフラインテストに劣らない結果が得られることを確認した．

Kai Akiyama[1,a)]　Yang Wu[1]　Kris M. Kitani[2]　László A. Jeni[2]

1. はじめに

人間の対面によるコミュニケーションにおいて，言語を用いた直接的な情報の交換だけでなく，表情や動作，視線などといった非言語情報は重要な役割を担っている．非言語情報のアノテーションは，コミュニケーションの心理学的な解析に使用されており，手作業で行われるアノテーションを自動化できれば解析の大幅な効率化が期待される．非言語情報のリアルタイムな認識は，ロボットが人間の感情を理解する手がかりとなり，また視覚障害者などのためのコミュニケーション支援システムへの応用が期待される．非言語情報のうち表情の認識については広く研究されている一方，うなずきや首振りといった頭部動作の認識に関する研究は数少ない．

頭部動作の自動認識を試みた従来の研究においては，画面の前で演じられた映像や，ロボットとのインタラクションの中で撮影された映像など，ヒューマンコミュニケーションとは異なる状況における頭部動作が研究対象となっていた．そのため，人間同士の会話における頭部動作の認識性能は明らかにされていない．また，扱われた頭部動作の種類も，うなずきおよび首振り程度の少数の種類に限定されていた．そこで本研究では，実際のヒューマンコミュニケーションにより近い自然会話映像に対して，多様な頭部動作を検出するという課題を提案する．この課題に取り組むにあたって，まず約 4 時間分の自然会話が収録され多種類の頭部動作がアノテーションされた，比較的大規模なデータセット (FIPCO) を構築する．また，一次元時系列データの変動を記述する HoVA 特徴量を提案し，その有効性を頭部動作解析において最先端のモデルである Latent-Dynamic Conditional Random Field (LDCRF) を用いた実験により確認する．さらに，LDCRF のオンラインテストへの適用可能性を示し，最後に，隠れ状態を用いて状態遷移をモデル化する LDCRF がサポートベクタマシン (SVM) よりこの問題に適していることを確かめる．

2. 関連研究

頭部動作検出に関連する研究は，大別して，データセットと検出手法の 2 つのグループに分けられる．この章ではこれらの研究について述べる．

2.1 データセット

頭部動作がアノテーションされた映像データセットは，3 つのグループに分けることができる．第一に，人間がコンピュータやロボットといった機械を相手にしたきの振る舞いが収集された，ヒューマンコンピュータインタラクションに関するものがある．

多くのデータセットでは，コンピュータを前にした被験者の振る舞いを固定カメラで撮影するという形がとられている．例えば，自発的な動作を収集するもの [5], [6], [13]，質問に Yes/No で答える際の動作を収集するもの [2], [10], [11]，およびその他のインタフェースによるもの [2] がある．し

1 奈良先端科学技術大学院大学
Nara Institute of Science and Technology
2 Carnegie Mellon University
a) akiyama.kai.ae6@is.naist.jp

ⓒ 2016 Information Processing Society of Japan

かし，これらには 2，3 種類程度の頭部動作が，種類あたり最大でも 2,3 百個程度しか含まれておらず，被験者の数も少ない．

ロボットとのインタラクションでは，コンピュータの画面を前にした振る舞いと比べると多少自然な頭部動作が収集できる可能性がある．[18] では，Mel と名付けられたペンギン型ロボットとのインタラクションを通して 2 つのデータセットが収集され，274 個のうなずきと 14 個の首振りが記録された．[12] では，ヒューマノイドロボットとの会話の中および事前に与えられた指示に基づく演技から，4 種類の頭部動作を数千個含む比較的大規模なデータセットが収集された．

こうしたヒューマンコンピュータインタラクションに関するデータセットには明らかな制約がある．まず，こうしたコンピュータを相手にした振る舞いは，人間を相手としたときと比べて強調されたり，慎重に振る舞われたり，あるいは繰り返されたりする傾向にあるため，撮影された頭部動作は不自然になりがちである．また，これらは一般に公開されておらず，収集した本人ら以外がベンチマークとして頭部動作手法を比較するために使用することができない．

第二に，言語学や感情科学の分野において収集されたデータセットがある．これらの分野では，人間同士の会話における自然発生的な動作が注目されているが，自然な動作の自動認識はより困難な課題となる．こうした分野におけるデータセットでは，あらゆる動作が研究対象となりアノテーションされる傾向にある．頭部動作がアノテーションされたデータセットの例として，聞き手の動作に注目した ALICO[4]，および相互方向の会話を記録した NOMCO[3] が挙げられる．これらはおよそ 10 種類の頭部動作が手動でアノテーションされているが，いずれについても映像から頭部動作の自動認識を試みた例はない．ただし，NOMCO のデータセットは異なる複数の環境で撮影されており，かつデータの総量も少ないことから，自動認識がより困難と思われる．ALICO のデータセットについては公開されておらず，自動認識に取り組む第三者がデータを活用することができない．

最後に，自然発生的な頭部動作の自動認識を目的として構築されたデータセットがある．話し手と聞き手の映像が撮影された RAPPORT データセット [7] は，仮想アバターが非言語情報によって人間との関係を築くために取るべき振る舞いを研究することを目的に構築された．この研究では頭部動作の理解と発生の予測に重点が置かれている．そこで，聞き手の頭部動作（うなずき）を含む非言語情報の予測を目的として，RAPPORT データセットに face-to-face サブセットが追加された [8], [9]．[15] および [16] では，2 者間の会話における頭部動作のミラーリング (行動の同調) を検出するために，ミラーリングを記録したデータセットが構築された．最近では，一対一の就職面接を撮影した Ubimpressed データセット，およびテーブルを囲む複数人での面接を撮影した KTH-Idiap コーパスが構築され，うなずき検出を行った研究がある [19]．しかしながら，これらのデータセットでは全てうなずきのみがアノテーションされており，うなずき以外の頭部動作については触れられていない．

本稿では，頭部動作の自動検出モデルを評価するために，後者 2 つのグループを参考にしたデータセットを構築する．そのために，人間同士の自然会話を収集し，多種類の自然発生的な頭部動作をアノテーションする．

2.2 頭部動作認識

頭部動作の検出には様々な手法が提案されてきた．典型的な手段としては，顔の中の目など目印となる部分を追跡し，映像中のフレーム毎に頭部姿勢を推定，スライディングウィンドウを用いた手法等によって頭部動作を検出するという方法がある．最も多く用いられる検出器は隠れマルコフモデル (Hidden Markov Model: HMM) である．[10] では，瞳孔の推定座標に基いて，うなずきおよび首振りを検出する 2 種類の離散型 HMM が使用された．この HMM の内部状態は事前定義された 3 つの状態であった．また，[11] もこれと同様の手法によるが，視線推定の手法が異なる．これらの研究ではいずれもヒューマンコンピュータインタラクションの状況を想定しており，Yes/No で答える質問によって誘発された頭部動作の映像が使用されている．その後，Fujie ら [12] は 4 つの連続型 HMM を用いて，それぞれうなずき，首振り，かしげ，静止状態を，その他の動作と区別する研究を行ったが，これはヒューマンロボットインタラクションを想定したものであった．近年の HMM の応用では，Wei ら [13] によるうなずきおよび首振りの検出，Terven ら [14] による頭部動作検出，ミラーリング動作検出 [15] が挙げられる．これらの HMM に基づくモデルにはいずれも，状態定義やルール定義を人の手に頼っているという共通した問題点がある．HMM の他には，SVM がよく用いられている．

HMM の他，スライディングウィンドウに基づく手法として，SVM が用いられることも多い [17]．頭部動作認識に関する初期の研究として，Morency ら [18] は SVM によるうなずきおよび首振りの検出を試みたものがある．近年では，Chen ら [19] が三次元顔追跡の出力から 2 種類の特徴量を抽出し，SVM によるうなずき検出を行っている．スライディングウィンドウのような抽出された特徴量を決められた時間幅のウィンドウ内で評価する方法は，SVM ばかりでなくランダムフォレストや CRF，HCRF といった他の識別器にも頭部動作検出アルゴリズムとして適用可能である [2], [16]．ただし，動作の長さが常に一定とは限らないことを考慮すると，スライディングウィンドウを適用

する際に通常はあらゆる幅のウィンドウでの探索を行うことが不可欠となる．また，SVM などにおいて各正事例の全体を切り出して学習する場合には，複雑な局所的変化を捉えることが難しいという問題もある．

これに対して，Morency ら [2] は HMM の変動をモデル化する能力を継承した LDCRF なる確率モデルを提案し，スライディングウィンドウにおける問題点を解決した．このモデルは HMM とは異なり，人の手によるルール定義を行わずに任意の数の隠れ状態を追加できる．また，スライディングウィンドウに頼ることなく時系列データ全体を一度に評価することが可能となっている．

LDCRF を含めた既存のモデルは非常に限られた頭部動作でのみ評価されてきた．特にその多くはうなずきのみを検出するものであり，多種類の自然な頭部動作の検出におけるパフォーマンスは明らかにされていない．本研究では独自に自然な頭部動作を収集したデータセットを使用して LDCRF を評価し，ベースラインとして SVM と比較する．さらに，LDCRF はオフラインテストだけでなくオンラインテストのためにも使用できることを確認する．

3. 手法

3.1 頭部動作検出の概要

頭部動作検出器構築の概要を図 1 に示す．多クラス検出へのアプローチとして，特定の頭部動作のみを検出できる独立した 2 クラス検出器を頭部動作の種類だけ用意する手法を採用する．これはアノテーションにあたって最大 3 種類までのラベルが重複することを許しているためである．テストの際には，各検出器へ同一のデータを入力し，各頭部動作ごとの性能を評価する．これにより複数のラベルが重複して検出されることを許して評価を行うことができる．実際のシステムとして使用する際は，最も高い尤度を出力した検出器のクラスを採用するなどの方法がある．

検出器を構築するために，まず学習用映像に対して頭部の位置姿勢推定を行う．ここでは三次元顔モデルフィッティング手法である Zface [20] を使用し，画像上における顔の X 座標，Y 座標，スケール，ピッチ角，ロール角，ヨー角を推定した．これら 6 変数から成る一次元時系列データから，後述する HoVA 特徴量を計算し，これとアノテーションされた真値の頭部動作ラベルを元に学習を行う．

3.2 HoVA 特徴量

Histograms of Oriented Gradients (HOG)[1] は，画像をセルに分割し，セルごとの勾配方向の分布をヒストグラムの形で表現した特徴量である．これには位置や大きさのばらつきに対してロバストであり，二次元の画像における物体認識では多くの支持を集めているが，1 次元の時系列データに対する応用例はない．しかし，動作認識においても位置や大きさのばらつきが存在するため，同様な特徴量

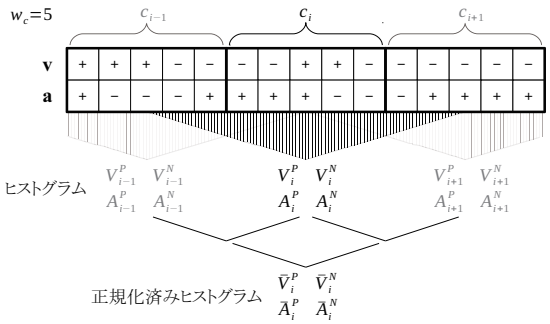

図 2 HoVA 特徴量

が 1 次元時系列データにも効果的である可能性がある．1 次元時系列データには勾配方向は存在しないが，勾配の正負が存在することに着目し，HOG 特徴量に類似した考え方が適用可能であると考えた．この 1 次元時系列データに適用可能な特徴量を Histogram of Velocity and Acceleration (HoVA) 特徴量と呼ぶこととする．

図 2 に従って HoVA 特徴量の詳細を述べる．まず，1 次元時系列データ $\mathbf{x}$ を一階微分した速度 $\mathbf{v} = \dot{\mathbf{x}}$ を求める．ここではマスク $[-1, 0, 1]$ を用いて一階微分を計算した．次に，時系列データを幅 w_c のセルに分割し，セルごとに正のビン V_i^P および負のビン V_i^N を用意する．これらのビンに，セルごとの $\mathbf{v}$ の符号の分布が投票される．このとき，$\mathbf{v}$ の注目セル c_i の中心から半幅 w_c 以内にある正・負の要素数を，注目セル c_i の中心からの距離による線形重みをかけて，それぞれ $V_i^P \cdot V_i^N$ に加える．最後に，こうして得られたヒストグラムを隣接するセルによって正規化する．ここでは，左側セルとの正規化と右側セルとの正規化の平均をとった．すなわち最終的な正規化された時系列データ $\bar{\mathbf{V}}^P$ および $\bar{\mathbf{V}}^N$ を次の式で与える．

$$\bar{V}_i^{P,N} = \frac{1}{2} \sum_{j \in \{0,1\}} \frac{V_i^{P,N}}{||V_i||_j} \tag{1}$$

$$||V_i||_j = \sqrt{(V_{i+j-1}^P)^2 + (V_{i+j-1}^N)^2 + (V_{i+j}^P)^2 + (V_{i+j}^N)^2} \tag{2}$$

以上を二階微分した加速度 $\mathbf{a} = \ddot{\mathbf{x}}$ に対しても行い，$\bar{\mathbf{A}}^P$ および $\bar{\mathbf{A}}^N$ を得る．二階微分のマスクは $[-1, 0, 2, 0, -1]$ を使用した．このようにして，時系列データ $\mathbf{x}$ から，長さ w_c のセルごとの 4 変数特徴量 $\bar{\mathbf{V}}^P, \bar{\mathbf{V}}^N, \bar{\mathbf{A}}^P, \bar{\mathbf{A}}^N$ を得る．頭部動作に適用する場合，頭部全体の位置姿勢の 6 変数それぞれに対して HoVA 特徴量が計算され，最終的に 24 変数となる．

HoVA 特徴量は，一次元時系列データの速度・加速度の変動を記述することができる．頭部動作は，頭部全体の位置姿勢の速度・加速度によって記述できると考えらえられ，また，速度・加速度を求める微分によって視点位置に依存しない検出が可能になる．HoVA 特徴量の利点として，セルに分割することでラベルのずれがある程度許容されるよ

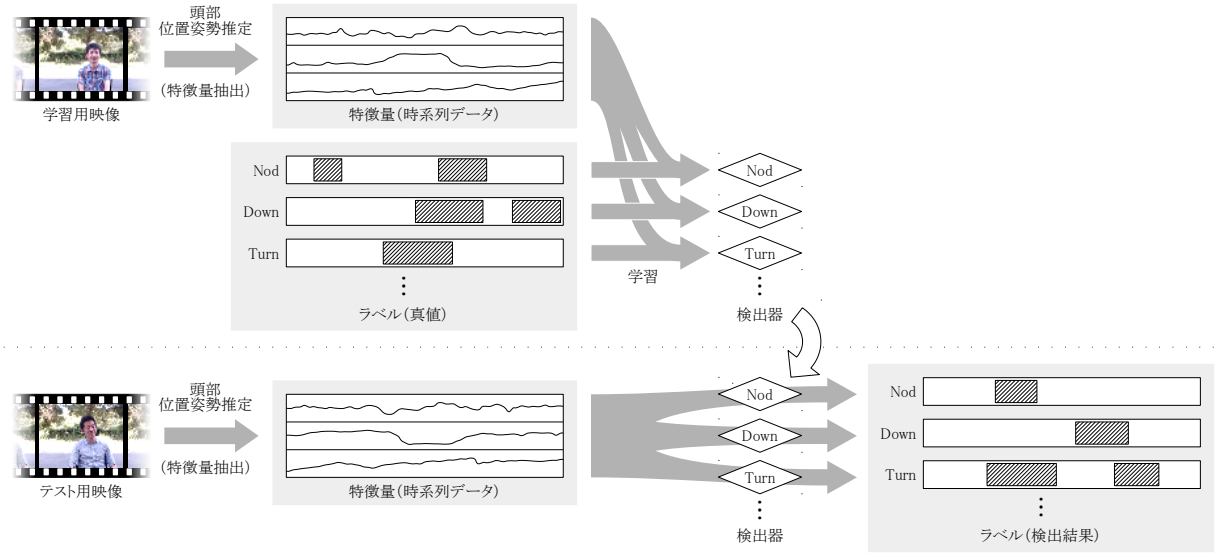

図 1 頭部動作検出器の学習とテストの枠組み

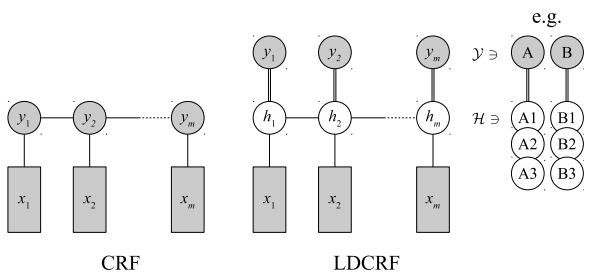

図 3 CRF と LDCRF のグラフ

うになる点，ヒストグラム化およびヒストグラムの正規化によってサンプル毎の波形の差異に対してロバストになる点が挙げられる．

3.3 LDCRF

Morency ら [2] は条件付き確率場 (Conditional Random Fields: CRF) を拡張し，LDCRF と呼ばれる無向グラフによる確率モデルを提案した．図 3 に示すように，CRF では教師データとして明示的に与えられるラベル間の状態遷移がモデル化される一方，LDCRF はそれに加えて隠れ状態層によりラベル内の暗示的な状態遷移をもモデル化することができる．

LDCRF では観測データ列 $\mathbf{x} = \{x_1, x_2, ..., x_m\}$ からそれぞれのデータに対応するラベル列 $\mathbf{y} = \{y_1, y_2, ..., y_m\}; y_j \in \mathcal{Y}$ を推定するためのモデルが学習される．なお，観測データ x_j からは何らかの特徴ベクトル $\phi(x_j)$ が計算される．それぞれが特定のラベルに紐付けられた隠れ状態 $\mathbf{h} = \{h_1, h_2, ..., h_m\}; h_j \in \mathcal{H}$ の導入によって，ラベル列の尤度は次の式で与えられる．

$$P(\mathbf{y}|\mathbf{x}, \theta) = \sum_{\mathbf{h}:\forall h_j \in \mathcal{H}_{y_j}} P(\mathbf{h}|\mathbf{x}, \theta) \tag{3}$$

$P(\mathbf{h}|\mathbf{x}, \theta)$ は CRF と同様な次の式を用いる．

$$P(\mathbf{h}|\mathbf{x}, \theta) = \frac{\exp(\sum_k \theta_k \cdot \mathbf{F}_k(\mathbf{h}, \mathbf{x}))}{\mathcal{Z}(\mathbf{x}, \theta)} \tag{4}$$

分配関数 $\mathcal{Z}$ は次の式で定義される．

$$\mathcal{Z}(\mathbf{x}, \theta) = \sum_{\mathbf{h}} \exp(\sum_k \theta_k \cdot \mathbf{F}_k(\mathbf{h}, \mathbf{x})) \tag{5}$$

$\mathbf{F}_k$ は，素性関数 f_k を系列の各時点で適用したものの総和である．

$$\mathbf{F}_k(\mathbf{h}, \mathbf{x}) = \sum_{j=1}^{m} f_k(h_{j-1}, h_j, \mathbf{x}, j) \tag{6}$$

素性関数 f_k には，観測素性および遷移素性がある．観測素性はそれぞれ特徴ベクトル $\phi(x_j)$ の特定の要素から特定の隠れ状態への対応を表す一方，遷移素性はそれぞれ特定の隠れ状態から隠れ状態への遷移を表す．従って，観測素性は $|\phi(x_j)| \cdot |\mathcal{H}|$ 個存在し，遷移素性は $|\mathcal{H}|^2$ 個存在する．

Morency らが用いている方法によると，学習用データは元のデータ系列から正事例と負事例が抽出されたものが与えられ，正事例については前後の負事例を含んだものが抽出される．テスト用データについては元のデータ系列がそのまま与えられ，系列全体で隠れ状態の遷移を最適化し，隠れ状態に対応するクラスのラベルが出力される．

3.4 スライディングウインドウによるテスト

スライディングウィンドウは，オンラインテストを行うための手段としてよく用いられている．SVM による検出ではテストにあたって一定の長さのデータしか受け付けないことから，マルチスケール探索付きのスライディングウィンドウを併用することになり，オンラインテストに適して

図 4　自然会話映像の撮影環境

いる．一方 LDCRF は，データが前後の隠れ状態の推測に影響を与えるため，Morency らはテストにあたって系列全体で最適化を行うオフラインテストを行った．しかし，リアルタイム性が要求されるシステムへの応用を前提とした場合，オンラインテストによる評価は不可欠である．そこで，スライディングウィンドウを LDCRF に導入し，オンラインテストを行った場合の結果を比較する．

なお，テストは HoVA 特徴量を用いて行うものとし，ウィンドウを HoVA 特徴量のセル単位で移動移動しながらテストする．スライディングウィンドウを導入するにあたっては，ラベル毎の尤度が同一の時刻に対して複数出力されるが，これらの平均を取ることで尤度を統合し，オフラインテストと同様に評価を行う．

4.　実験

HoVA 特徴量の有効性，LDCRF にスライディングウィンドウを導入した場合の影響，および LDCRF の SVM に対する優位性を検証するために，異なる特徴量またはモデルによって学習を行い，テストによって評価する実験を行う．以下ではまず実験のために収集されたデータセットについて述べ，実験環境について説明した後，各実験の結果を示す．

4.1　データセット

実験に用いるデータセットとして，自然会話映像が撮影されたものを使用した．会話の話題は事前に提供されるリストから被験者同士の相談により決定され，与えられた対立する 2 つの物事についての議論，および与えられた課題に対して計画を立てる議論の両方を含んでいる．計 15 名の被験者から毎回異なる組み合わせの 2 名が参加し，各映像約 11 分，全 30 本の会話映像が収集された．図 4 のように，撮影にはウェアラブルカメラおよび固定カメラが使用され，約 24fps で記録された．本稿では固定カメラによって撮影されたものを実験に用いた．

撮影された映像に対して，10 種類の頭部動作 (表 1) が 3 種類まで重複を許してアノテーションされた．なお，アノテーションする頭部動作の種類については [3] および [4] を参考に決定した．アノテーションされた頭部動作の内訳は図 5 に示す通り，うなずきが最も多い割合を占めている．

また，Zface による頭部位置姿勢推定が行われ，X，Y，スケール（頭部の大きさ），ピッチ，ロール，ヨーからから成る 6 種類の時系列データを得た．その後，ノイズを軽減するために 3 フレーム幅で平滑化処理を行った．

4.2　実験環境

実験方法は，交差検証を除いて Morency らによる方法に従う．すなわち，学習では正事例を抽出したデータと負事例を抽出したデータを使用する．正事例を抽出する際には，前後の 2〜50 フレームに相当するデータを含めて切り出すことで，正負のラベル間の遷移が学習されるようにする．負事例を抽出する際には，30〜60 フレームの長さに相当する純粋な負事例の領域をランダムに切り出す．Morency らの手法とは異なる点として，各学習での交差検証は行わず，ラベルあたりの隠れ状態数として 2 または 5，L2 正則化項のパラメータとして 25 を採用した．また，最適化の繰り返し回数は 100 回とした．なお，LDCRF による学習およびテストは hCRF Library[*1] による．

本稿では leave-one-sequence-out (LOSO) による評価を採用する．すなわち，30 本の全映像の内 29 本の映像で学習を行い，残りの 1 本の映像でテストを行う．このセットを 1 スプリットとし，テストに用いる映像を変えながら 30 スプリットにわたって繰り返した後，全てのスプリットで得られた結果を総合して評価する．

評価の基準として，PR 曲線および平均適合率 (average precision: AP) を用いる．PR 曲線は，検出に用いる尤度の閾値を変化させることで再現率 (recall) が 0 から 1 まで変化した際の適合率 (precision) の変化を示した曲線である．平均適合率は，各再現率における平均の適合率である．

4.3　HoVA 特徴量の有効性

4.3.1　位置姿勢をそのまま用いた場合

位置姿勢の 6 チャンネルの時系列データをそのまま特徴量として用いて実験を行った．LDCRF におけるウィンドウ幅は 13 フレーム, ラベルあたり隠れ状態数は 2 および 5 が使用された．データ量が非常に多く，学習に多くの時間を必要とするため，比較的サンプル数の豊富な Nod および Turn のみの結果の PR 曲線および平均適合率を図 6 に示す．いずれの場合も，正または負のどちらか片方に極端に偏った結果が得られた．従って，PR 曲線はランダムに推定を行った場合に近い曲線となった．自然会話における頭部動作にはアノテーションされない細かな動きが多く，うなずきひとつをとっても強度の違いなど多数のモードが存在する．また，Zface による位置姿勢推定における誤差がノイズとして位置姿勢データに現れることもあり，モデルがデータに適合することが非常に困難であったと予想される．

*1　https://sourceforge.net/projects/hcrf/

ⓒ 2016 Information Processing Society of Japan

主な移動・回転軸	ラベル名 (英)	ラベル名 (和)	説明
ピッチ	Nod	うなずき	頭が下がって上がる
	Jerk	上方うなずき	頭が上がって下がる
	Up	仰ぎ	頭がしばらく上げ続ける
	Down	俯き	頭がしばらく下げ続ける
	Ticks	連続うなずき	連続したうなずき
ロール	Tilt	かしげ	片方に頭を傾ける
ヨー	Shake	首振り	水平に頭を降る
	Turn	横を向く	水平に頭を回転させる
Z (スケール)	Forward	乗り出し	前に乗り出す
	Backward	退き	後ろに退く

表 1 頭部動作の定義

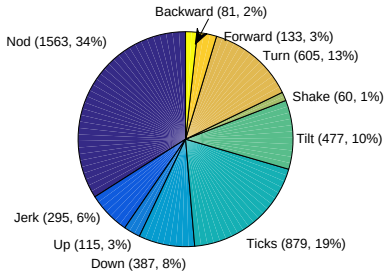

図 5 ラベルの内訳．括弧内は数と全体に占める割合を表している

4.3.2 HoVA 特徴量を用いた場合

位置姿勢の 6 チャンネルの時系列データを各チャンネルごとに HoVA 特徴量に変換し，これを用いて実験を行った．LDCRF におけるウィンドウ幅は 3 セル (=15 フレーム)，ラベルあたり隠れ状態数は 2 および 5 が使用された．

前節の実験と比較するため，Nod および Turn のみの結果の PR 曲線および平均適合率を図 7 に示す．位置姿勢を用いた場合と比較するとより高い平均適合率となり，HoVA 特徴量の有効性が確認できる．

全ての頭部動作に対するテストの結果を図 8 に示す．平均適合率が最も低い Shake は，サンプル数が 60 個と最も少なく，学習するには少なすぎたことは明らかである．

4.4 スライディングウインドウによるテスト

HoVA 特徴量を用いて学習されたモデルを用いて，スライディングウインドウによるテストを行った．LOSO の各スプリットごとに，ウィンドウ幅として学習データの正事例の平均長さが用いられた．全ての頭部動作に対するテストの結果を図 9 に示す．図 8 と比較すると Nod など性能が低下しているクラスもあるが，Down など逆に向上しているクラスも見られる．概して，スライディングウィンドウによるテストは Morency らの手法と同様に有効であると言える．

4.5 SVM との比較

頭部位置姿勢特徴量および HoVA 特徴量を用いて SVM による学習とテストを行った結果を図 10 に示す．頭部位置姿勢特徴量を用いた場合は，LDCRF と同じく，正か負のどちらかに極端に偏った検出結果が得られた．従って，図 6 と同様な PR 曲線を描いている．次に，HoVA 特徴量を用いた場合は，頭部位置姿勢特徴量を用いた場合より良好であるものの LDCRF(図 8) には及ばない結果となった．よって，この課題に関しては，LDCRF は SVM よりも優

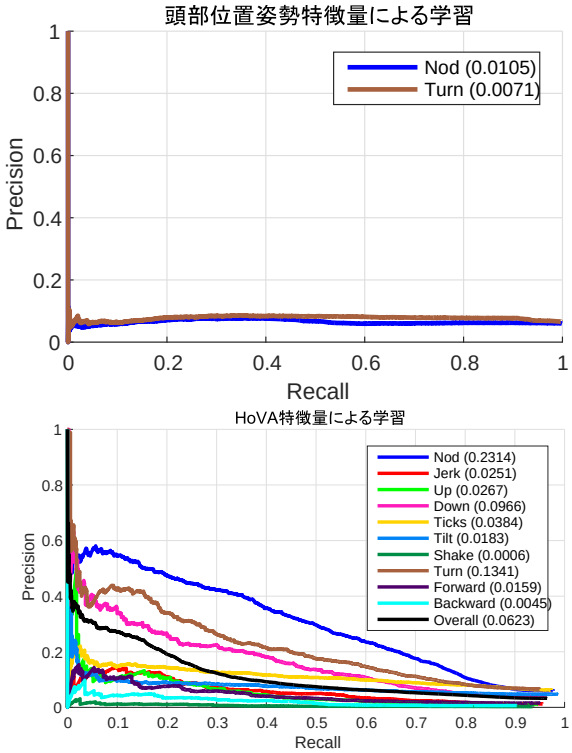

図 10 SVM によるテスト結果の PR 曲線

れていると言える．

LDCRF が SVM に比べて優れている主な理由の一つとして，学習とテストの方法の違いが挙げられる．SVM におけるテストでは，ウィンドウが純粋な正事例か負事例かを識別するため，正負のラベル間の遷移を考慮しないが，LDCRF ではラベル間の遷移を学習し，検出に資することができる．

その他の理由として，モデルが持つパラメータ数の柔軟性の違いが挙げられる．SVM では，モデルのパラメータの数は特徴ベクトルの次元やウィンドウサイズによって決められる．一方 LDCRF では，隠れ状態の数を調節することで，特徴ベクトルやウィンドウサイズを変更することな

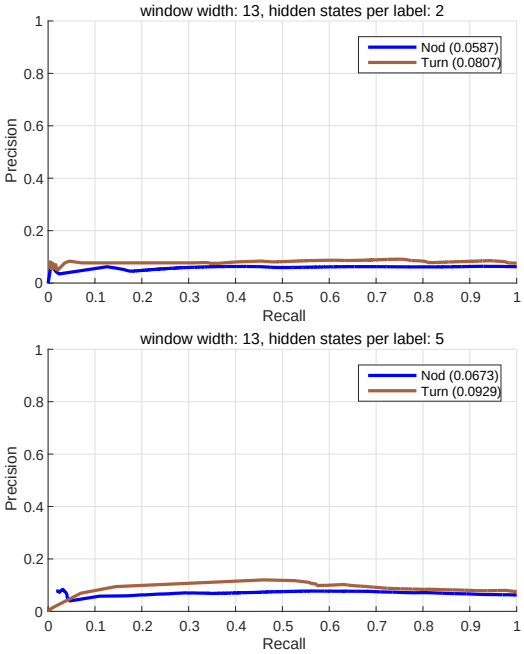

図 6　位置姿勢特徴量を用いた
Nod,Turn のテスト結果の PR 曲線

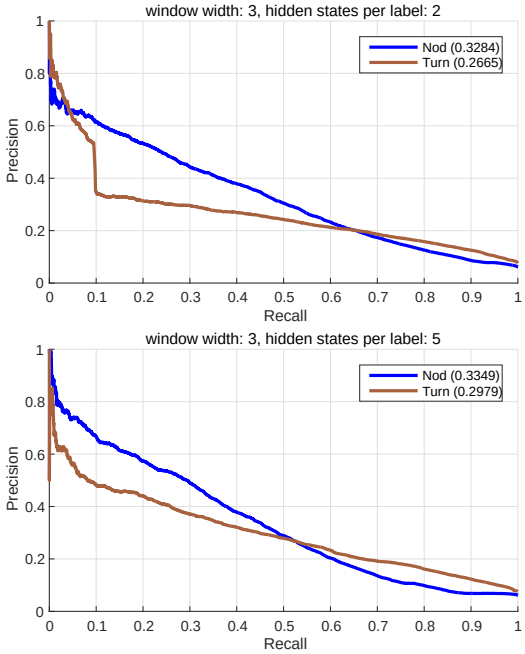

図 7　HoVA 特徴量を用いた
Nod,Turn のテスト結果の PR 曲線

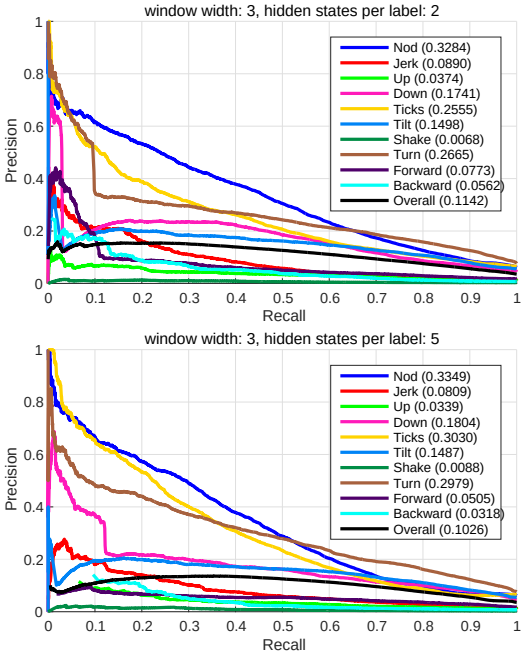

図 8　HoVA 特徴量を用いたテスト結果の PR 曲線

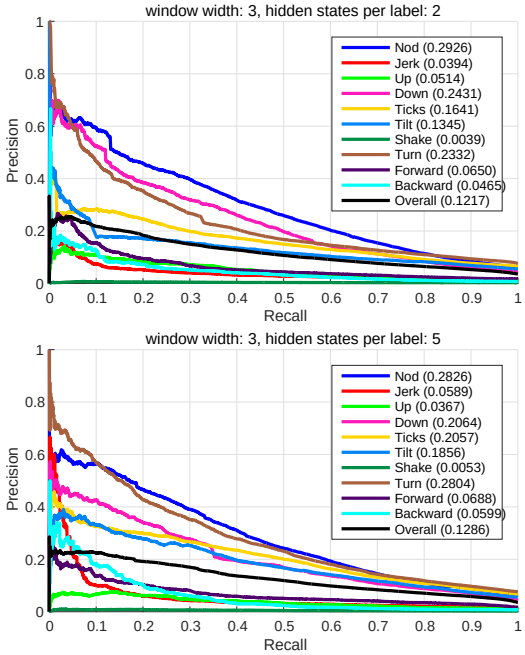

図 9　スライディングウィンドウによるテスト結果の PR 曲線

ⓒ 2016 Information Processing Society of Japan

くモデルのパラメータの数を調節できる．すなわち，より複雑なデータに対してはより多くの隠れ状態を使用することでモデルが適合しやすくなると考えられる．

5. おわりに

本研究では，自然会話における多種類の頭部動作を自動検出することを目的として，HoVA 特徴量を提案し，その有効性を実験によって検証した．また，LDCRF はスライディングウィンドウの導入によってオンラインテストが可能であること，および SVM に対する LDCRF の優位性を実験によって確認した．HoVA 特徴量は，FIPCO データセットにおける頭部動作検出での有効性が示されたが，他のデータセットに対しても有効に作用する可能性がある．また，LDCRF にスライディングウィンドウを導入し，オンラインテストを行うことで，性能を大きく損なうことなくリアルタイムで高速に検出を行うことが可能である．

今後の課題として，本稿で取り上げた手法の頭部動作検出ではなく識別への適用，および固定カメラではなくウェアラブルカメラの映像による頭部動作検出が挙げられる．

謝辞　本研究は JSPS 科研費 15K16024 の助成を受けたものです．

参考文献

[1] Dalal, N. and Triggs, B.: Histograms of oriented gradients for human detection, 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), pp. 886–893 (2005)

[2] Morency, L.P., Quattoni, A. and Darrell, T.: Latent-dynamic discriminative models for continuous gesture recognition, 2007 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8 (2007)

[3] Paggio, P., Allwood, J., Ahlsn, E., Jokinen, K. and Navarretta, K.: The NOMCO multimodal Nordic resource - goals and characteristics, Proceedings of the 7th Language Resources and Evaluation Conference, pp. 2968–2973 (2010)

[4] Buschmeier, H., Malisz, Z., Skubisz, J., Wlodarczak, M., Wachsmuth, I., Kopp, S. and Wagner, P.: ALICO: A multimodal corpus for the study of active listening, Proceedings of the 9th Language Resources and Evaluation Conference, pp. 3638–3643 (2014)

[5] Choi, H.-I. and Rhee, P.-K.: Head gesture recognition using HMMs, Expert Syst. Appl. 17 (3), pp. 213–221 (1999)

[6] Kawato, S. and Ohya, J.: Real-time detection of nodding and head-shaking by directly detecting and tracking the "between-eyes", Proc. Fourth IEEE International Conference on Automatic Face and Gesture Recognition, pp. 40–45 (2000)

[7] Gratch, J., Wang, N., Gerten, J., Fast, E. and Duffy, R.: Creating Rapport with Virtual Agents, 7th International Conference on Intelligent Virtual Agents, pp. 125–138 (2007)

[8] Morency, L.P., De Kok, I. and Gratch, J.: Predicting Listener Backchannels: A Probabilistic Multimodal Approach, Proc. 8th International Conference on Intelligent Virtual Agents pp. 176–190 (2008)

[9] Ozkan, D. and Morency, L.P.: Latent Mixture of Discriminative Experts, IEEE Transactions on Multimedia 15 (2), pp. 326–338 (2013)

[10] Kapoor, A. and Picard, R.W.: A real-time head nod and shake detector, Proc. ACM Int'l Conf. Perceptive User Interfaces, pp. 1–5 (2001)

[11] Tan, W. and Rong, G.: A real-time head nod and shake detector using HMMs, Expert Syst. Appl. 25 (3), pp. 461–466 (2003)

[12] Fujie, S., Ejiri, Y., Nakajima, K., Matsusaka, Y. and Kobayashi, T.: A conversation robot using head gestur, e recognition as para-linguistic information, Proc. IEEE Int'l Workshop Robot and Human Interactive Communication, pp. 159–164 (2004)

[13] Wei, H., Scanlon, P., Li, Y., Monaghan, D. and O'Connor, N. E.: Real-time head nod and shake detection for continuous human affect recognition, Image Analysis for Multimedia Interactive Services (WIAMIS), pp. 1–4 (2013)

[14] Terven, J.R., Salas, J. and Raducanu, B.: Robust head gesture recognition for assistive technology, Volume 8495 of the series Lecture Notes in Computer Science, pp. 152–161 (2014)

[15] Terven, J.R., Raducanu, B., Meza-de-Luna, M.E. and Salas, J.: Head-gestures mirroring detection in dyadic social interactions with computer vision-based wearable devices, Neurocomputing 175, pp. 866–876 (2016)

[16] Terven, J.R., Raducanu, B., Meza-de-Luna, M.E. and Salas, J.: Evaluating real-time mirroring of head gestures using smart glasses, The IEEE International Conference on Computer Vision (ICCV) Workshops, pp. 452–460 (2015)

[17] Huang, D., Yao, S., Wang, Y. and De la Torre, F.: Sequential Max-Margin Event Detectors, Proc. The 13th European Conference on Computer Vision, Part III, pp. 410–424 (2015)

[18] Morency, L.P., Sidner, C., Lee, C. and Darrell, T.: Contextual recognition of head gestures, Proceedings of the 7th international conference on Multimodal interfaces (ICMI '05), pp. 18–24 (2005)

[19] Chen, Y., Yu, Y. and Odobez, J.M.: Head nod detection from a full 3d model, The IEEE International Conference on Computer Vision (ICCV) Workshops, pp. 136–144 (2015)

[20] Jeni, L.-A., Cohn, J.F. and Kanade, T.: Dense 3D face alignment from 2D videos in real-time, Proc. 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), pp. 1–8 (2015)

ⓒ 2016 Information Processing Society of Japan