コンピュータビジョンにおける時系列パターン認識

2006-CVIM-154 p197–209（産総研の西村さんと共著） の川嶋担当部（3 節:モデルベース, 4 節:応用分野）．ただし，3 節の一部（VLMM まわり）修正.

Temporal Pattern Recognition in Computer Vision

3. モデルに基づく時系列パターン認識

軌道のマッチングに基づいて認識を行う場合，わずかなデータでも高速な学習が実現できる反面，時系列パターンの分布を表現できない．そこで，1990 年代ごろからは，時系列パターンの変化に関して，あらかじめ状態遷移に基づくモデルを持つことで認識を行う，モデルベースの手法が多く用いられるようになっている．

時系列パターンの認識やモデル化に関しては，ビジョンに限らず，音声や言語，機械学習，制御，統計，生物学（DNA 解析）といった多くの分野での興味の対象であり，ビジョンの分野における時系列パターンのモデルも，音声や言語などの分野で行われてきたものを，ビジョン特有の特徴量や表現を導入・拡張した上で利用している場合が多い．特に，hidden Markov model (HMM)[6)~8)] は，現在の音声認識の標準的な認識手法として用いられており，ビジョンの研究でも最も多く用いられている時系列パターンのモデルである．一方で，ジェスチャや表情などの人の動きを認識・解析する上では，その速度や持続時間といった力学的な特徴が重要になる場合も多い．そこで，制御のシステム方程式や微分方程式といった力学的なモデルが，1990 年代後半からしばしば用いられている．

どのモデルを用いるかは，認識したい対象の時間的振る舞いによって使い分ける必要があり，本稿では以下のような分類をする☆.

(1) 離散事象系 (discrete event system)
(2) 力学系 (dynamical system)
(3) 両者の混在系 (hybrid (dynamical) system)

離散事象系を対象とする場合は，離散状態の集合をまず定義しておき，離散状態の順序構造をモデル化する．これには，有限状態機械に基づく方法[9)] や，その確率的モデルの 1 つである HMM[10)] があり，さらに複雑な構造を扱うために確率化した文脈自由文法もしばしば用いられる[11)]．離散事象系のモデルは,「トランプカードを切る」,「持ち上げる」といった比較的抽象度の高い行動の間の関係（例えば文脈を踏まえた行動解析など）を表現するのに適しているが，HMM に関しては「足を上げる」際の角度の変化といった，信号レベルに近い変化のモデル化にも多く用いられる．

力学系を対象とする場合は，連続状態空間（実ベクトル空間など）を仮定し，微分方程式や差分方程式によって状態変化を記述する．ビジョンの分野では，制御理論の線形動的システムや，自己回帰モデル，パーティクルフィルタが多く用いられている．HMM などの離散事象系が連続的な信号変化（特徴量変化）の情報を失うのに対して，力学系モデルは連続的な変化をそのままダイナミクスとして表現できるため，認識だけでなく生成にもしばしば用いられる[12)]．ただし，線形な力学系では複雑な構造を扱うことが困難であるため，離散事象系と力学系を統合したハイブリッドシステムが，特に 1990 年代後半から，ビジョンや制御などの多くの分野で提案されている．

以下では，離散事象系，力学系，ハイブリッドシステムの 3 つの分類に従って，3.1 節，3.2 節，3.3 節においてそれぞれの代表的な手法を紹介する．

3.1 離散事象系に基づく時系列パターン認識

3.1.1 有限状態機械を用いる手法

有限状態機械（finite state machine, FSM）（もしくは有限オートマトン）では，HMM と異なり，状態が直接観測できるものと考える．認識したい事象が離散的な状態の系列としてあらかじめ記述できる場合（事象集合や要素間の順序，分岐構造などが既知，シナリオが与えられている場合），もしくはその構造が学習可能な場合にしばしば用いられる．

☆ 系（システム）は実際の対象（実体）を指し，モデルは数式などによるシステムの記述方法を指すが，系（システム）という用語はモデルとしての意味で用いられることも多い.

Bobick らの提案したジェスチャ認識手法[9] では，学習時には，手やマウスのジェスチャから得られた特徴空間中のサンプル点集合から，個々のクラスの典型的な曲線 (prototype curve) を見つける．次に，この prototype curve を離散化し，得られたベクトル集合をクラスタリングすることで状態の集合を決定する．このとき，状態は複数のクラス間で共通して用いられる場合もある．すると，各クラスのジェスチャは，この状態集合中の状態の系列としてそれぞれ表現できるようになる．認識時には，得られた特徴点の系列を状態系列にした上で，学習時に得られている各クラスの典型的な状態系列と，動的計画法によってマッチングを行うことでクラス識別をする．

日常の生活で現れるより自然なジェスチャを表現するために，同じく Bobick らは，各状態の持続長分布や動きの大きさの分布を持たせたモデルを提案している[13]．ジェスチャにおける動きの休止の入れ方などに基づいて，二相性 (bi-phasic) と三相性 (tri-phasic) のジェスチャを認識している．なお，このとき持続長をモデル化した HMM[14] を参考にしているが，これについては 3.1.2 節で述べる．

Wada らは，非決定性有限オートマトンを用いることで，複数人物の入退室の認識を行った[15]．このとき，各視点のカメラについてそれぞれオートマトンを設け，認識時には互いの状態に対して抑制をかけることで，動的に仮説を統合する機構を導入している．

3.1.2 Hidden Markov Model を用いる手法

ビジョンにおける HMM の利用としては，手書き文字の認識が 1980 年代中ごろより行われているが，特にカメラで撮影した動画像を入力として時系列パターン認識を行ったものとしては，Yamato らのテニスのスイングの認識[10),16)] が初期のものであり，さらに Starner と Pentland による手話認識[17] がよく知られている．その後，HMM はジェスチャや表情，行動認識などの様々な目的に用いられ，様々な拡張が提案されているが，ここでは特に

- 空間的変動のモデル化 (a)
- 時間的変動のモデル化 (b)
- 複雑な状態遷移の構造のモデル化 (c,d)
- 複雑な依存関係のモデル化 (e)
- HMM の集合（クラス）の学習法 (f)

のそれぞれに焦点を当てながら分類を行う（括弧内はパラグラフ番号）．

HMM は状態遷移によって時間方向の伸縮に対応し，各状態での出力分布によって空間的な変動に対応するが，以下で述べるように，目的に応じてこれらにさらに拡張が行われる．特に時間方向の変動に関しては，状態の持続長を明示的に持つマルコフモデルとして segment model[18] が提案されている．これらは HMM と区別されることもあるが，本稿では HMM と共にまとめて紹介をする．

HMM の状態遷移のトポロジー（状態の接続関係に関する構造）として，構造や状態数の学習方法についての検討がなされている．また，階層的な構造を持った HMM がいくつか提案されており，これらについても紹介をする．

複数のプロセス間の相互関係（モダリティ間の関係や左右の手の動きの関係など）や，過去への依存関係を柔軟に表現するために，状態変数や観測間の様々な依存関係（因果関係など）をモデル化した HMM について紹介を行う．これらを統一的に記述する dynamic Bayesian network については，3.1.3 節でまとめて紹介を行う．

なお，各状態が出力だけでなく入力も持つモデルとして input-output HMM があり，ジェスチャ認識への応用が検討されているが[19),20)]，ほとんどの場合には，HMM を生成モデルとして用いる．したがって，以下ではデータを HMM へ「入力する」という言い方をした場合は，その HMM の観測データとして用いることを意味するものとする．

(a) 空間的変動をモデル化するための HMM

HMM では出力確率の分布によって状態と特徴量の対応をモデル化するため，どのような分布を用いるかが，特徴選択と共に重要となる．ビジョンにおける初期の認識手法としては，音声認識と同様に，離散 HMM (discrete HMM) を用いた手法[10),16),17)] や連続分布 HMM (continuous density HMM) による手法[21),22)] が用いられ，現在も多くの認識システムに利用されている．一方で，ジェスチャなどの認識においては音声以上に特徴量の変動が大きくなるという問題があり，これを解決するために，出力確率に対してパラメタを設け，系統的な変動をモデル化した parametric HMM が提案されている[23),24)]．最近では，出力確率をパラメトリックな分布としてではなく事例として持つという考え方による手法[25] があり，これは非常に多くの出力記号を持つ離散 HMM とも考えることができる．

(b) 持続長をモデル化するための HMM

HMM はもともと離散事象系の確率モデルであり，状態遷移の時間的なタイミングは表現しない（オートマトンでは基本的には入力が得られたときに状態遷移を行う）．特に，通常の認識システムでは固定長の周期でサンプリングした特徴ベクトル系列を入力として用いるため，各サンプリング時刻で状態遷移を考えることになる．ある離散時刻における HMM の状態を i とし，次の時刻で再び同じ状態に遷移する確率を a_{ii} とすれば，持続時間 t（t サンプリング）だけ状態 i が持続する確率は，t 回セルフループする確率として計算できるため

$$d_i(t) = a_{ii}^t(1 - a_{ii}) \tag{1}$$

となる[7]．$d_i(t)$ は持続時間が長くなるほど確率が低くなる指数分布となるため，特定の持続時間の確率が高い場合などは，適切にモデル化できない．そこで，

音声の分野では，各状態の持続長分布を明示的に持たせた time duration HMM[6),7),14)] が提案されている．これをさらに持続時間を持ったマルコフモデルとして一般化した segment model[18)] もしくは hidden semi-Markov model が提案されている．また，マルチグラムモデル[26)] と呼ばれるモデルが，入力系列を可変長の系列に分節化する際に用いられることがある．time-duration HMM, segment model およびマルチグラムモデルの関係も含めて，Murphy によって統一的に述べられている[27)]．Segment model の中には，自己回帰モデルや動的システムと組み合わせたものも含まれているが，これらは 3.3 節にて改めて述べる．

ビジョンの分野では，3.1.1 節で述べた文献[13)] 以外にも，time duration HMM に基づく時系列パターン認識がしばしば検討されている[28),29)]．

(c) HMM のトポロジーの学習法

HMM の状態間の接続関係は経験的に決めるられることが多く，特に left-to-right モデルがよく用いられる．ただし，認識対象によっては最適なトポロジーが自明でない場合があり，自然言語の分野では，Brants によって状態のトポロジーを学習する手法が提案されている．これには，状態のマージに基づく手法[30),31)] と状態の分割に基づく手法[32)] がある．ビジョンの分野では，これらの手法とは独立に，HMM の状態のクラスタリングを EM アルゴリズムに基づいて行う方法[33)] や，エントロピーの最小化によって HMM の状態数やそのトポロジーを学習する手法[34)] が提案されている．

(d) 階層的な構造を扱うための HMM

複雑な事象は，簡単な事象の組み合わせとして表現できることが多く，HMM に階層的な構造を導入した hierarchical HMM が Fine らによって提案されている[35)]．これは，HMM の各状態が出力確率ではなく，下位の HMM を持てるようになっている．そして下位の HMM の状態もさらにその下位の HMM を持てるという再帰構造になっている．いったん下位の HMM のいずれかの状態に遷移した後に，その HMM の最終状態に遷移すると，再び元の上位の状態に戻る．したがって再帰的に下位の状態を活性化させることができ，後述する確率文脈自由文法 (SCFG) の簡略化されたモデルとなっている．また，SCFG よりも少ない計算量で学習することができる．Hierarchical HMM の拡張[36)] やビジョンへの応用[37)] は Bui らによって積極的に行われており，室内での行動解析などに用いられている．

Oliver らによる layered HMM[38)] では，hierarchical HMM とは異なる考え方で階層化を行っている．これは，下位の HMM 組（クラス数を K とする）の尤度（K 次元ベクトル，もしくは最大の尤度）を，上位の HMM への入力とするものであり，オフィス環境におけるマルチモダリティを統合した行動認識システムへ応用されている．

(e) 複数のプロセス間の関係を扱うための HMM

両手の動きからなるジェスチャは，右手と左手の時系列データの組からなると考えることができる．このような複数の時系列データを認識するために，複数の HMM を結合した coupled HMM と呼ばれるモデルが Brand によって提案されている[41)]．2 つの HMM を結合する場合を例にとれば，一方の HMM のある時刻における状態に対する，他方の HMM の次の時刻の状態を確率として持つことで，2 つの HMM の状態の共起関係をモデル化することができる．

(f) 複数の HMM のクラスタリング

一般的な HMM による時系列パターン認識では，クラスの数だけ HMM を用意しておき，入力された時系列に対して最も高い尤度を持つ HMM を決定することでクラス識別する．しかし，そもそも何種類の HMM が必要であるかが未知である場合も多く，さらに個々の HMM を学習するためには，あらかじめクラスごとに分節化された学習データが必要になるという問題がある．これらの問題を，複数の HMM の（モデルベースの）クラスタリングによって解決しようとする方法が検討されている[42),43)]．

3.1.3 Dynamic Bayesian Network

ベイジアンネットの観点から，HMM における状態をいくつかの因子が組み合わさったものとして表現できるようにしたものを，dynamic Bayesian network (DBN, もしくは dynamic belief network) と呼ぶ．Ghahramani は，HMM と離散時間カルマンフィルタとの共通性について DBN の観点からの説明している．特に，HMM の前向きアルゴリズムとカルマンフィルタリング，HMM の後向きアルゴリズムとカルマンスムージング，さらに両者の学習方法について，DBN の確率推論・学習の観点から統一的に述べている[44)]．また，Ghahramani 自身の提案している factorial HMM[45)] についても DBN の観点から解説がなされている．

Murphy は，前節で述べた様々な HMM について DBN の観点からまとめると共に，DBN の確率推論の近似方法などについても検討している[46)]．DBN はこの数年で急速に広まっており，特に行動解析・認識において複雑な文脈を表現する場合や，複数の時系列データの統合などの，構造を持った時系列パターンの認識に用いられることが多い[47)~49)]．

3.1.4 2 次以上のマルコフ性のモデル化

HMM や DBN は，通常単純マルコフ性を仮定して，離散時刻 t の状態 s_t が，直前の時刻 $t-1$ の状態 s_{t-1} によってのみ決まるものとしている．ただし，状態によっては直前だけでなくさらに過去の状態に依存する場合がある．そこで，2 次以上のマルコフ性を導入することが考えられる[8)]．ただし，どれだけ過去に依存するかが状態によって異なる場合がある．そこで，機

械学習の分野では可変長 N グラムモデルと呼ばれるモデルが提案されており[39]，マルコフ性の次数を状態に応じて可変となるようなマルコフモデルを学習できる．これに基づいて，ビジョンの分野では Galata らによって variable-length Markov model (VLMM) を用いる方法[40] が提案され，エクササイズにおける人物の輪郭変化パターンを学習・生成することで，次数を上げることの有効性を検証している．

3.1.5 文脈自由文法に基づく手法

有限状態機械 (FSM) と HMM の関係と同様に，文脈自由文法の確率モデルとして確率文脈自由文法 (stochastic context-free grammar, SCFG) があり，構造的な事象を，HMM よりも直接的に扱うことが可能である．Ivanov と Bobick は，SCFG をジェスチャ認識および駐車場における状況認識に適用している[50]．これは，プリミティブとなる離散事象（手を上げる，下げるといった単純な動きなど）を入力された動画像から検出していく（例えばジェスチャ認識では HMM を利用）．あらかじめこれら事象間の関係（文脈）を SCFG によってモデル化しておくことで，確率推論によってあいまい性を解消しながら安定した認識が可能となる．

Moore と Essa は，Ivanov らの手法に対して誤り検出やリカバリーなどの拡張を行い，カードゲームにおける各プレイヤーの行動 (behavior) や戦略の認識を行っている[11]．

3.2 力学系に基づく時系列パターン認識

力学系モデルは時間の経過と共に自律的に状態を遷移させていくことのできるモデルであり，時系列パターンの認識や表現に用いられる線形な力学系モデルとしてはカルマンフィルタなど（線形動的システム）が，非線形な力学系モデルとしては，拡張カルマンフィルタ，リカレントニューラルネットなど（非線形動的システム）などがよく用いられる☆．離散時間の動的システムでは，あるサンプリング時刻の状態（実ベクトル）が，直前のサンプリング時刻の状態にのみ依存するという単純マルコフ性を仮定する．また，状態に対して変換を行うことで，観測データ（実ベクトル）が得られる．このように，単純マルコフ性を仮定した状態遷移，および状態からの観測によって時系列をモデル化する点で，動的システムと HMM は類似点が多い（3.1.3 節や文献[51] を参照のこと）．

しかし，動的システムが時間の経過にしたがって自律的に状態を遷移させる（離散時間モデルはその近似である）のに対し，HMM は状態の順序のみをモデル化し，入力が得られるたびに状態遷移を考える（離散事象系である）という点で，両者の時間や状態の考え方は全く異なる．そのため，動的システムを時系列パターン認識へ用いる際も，状態（ジェスチャのポーズや離散的事象）の順序というよりは，その変化の度合いや速度，力のかかり方といった物理的特性そのもの（メトリックを持った情報）が重要になる場合に用いられる．また，動画像や関節の動きといった時系列パターンの生成にしばしば応用されている[12),52]．

3.2.1 線形動的システム

線形動的システムとしては，状態に単一のガウス分布を仮定するカルマンフィルタがよく知られている[53]．一方，複数人物のトラッキングなどではこの仮定が成り立たない場合も多い．そこで，粒子（パーティクル）の集まりによって状態の非ガウス分布を表現する方法として，Blake らの CONDENSATION[54] を始めとするパーティクルフィルタがある．

状態にガウス分布を仮定した線形動的システムを用いて，時系列パターンの認識を行う初期の研究としては Bregler[55] のモデルがあるが，これは HMM との混在系の中で用いており，3.3 節で改めて述べる．線形動的システムを単体で用いるものとして，Rao は，ロバスト推定とカルマンフィルタを組み合わせることで，遮蔽物があるような入力画像に対して，あらかじめ学習された画像系列を想起できる手法を提案している[56]．これに近いモデルとして，与えられた動画像から線形動的システムを同定し，同定されたシステムから動画像の生成や認識を行う方法（Dynamic texture）を提案している[12]．Rittscher と Blake は，自己回帰モデルに基づいてエアロビクスにおける人の動きをモデル化し，システム行列の固有値などを解析している[57]．Bissacco らは，人の walking,running,dancing,jumping などの動きをそれぞれ線形動的システムとしてモデル化し，あらかじめシステム間の距離を定義しておくことで，認識時には，登録されたシステムと入力データから同定されたシステムとの距離による nearest neighbor によって認識を行っている[58]．

状態の非ガウス分布を導入した手法として，Blake らは，3.3 節で述べるハイブリッドシステムの観点から，ジャグリングにおける玉の動きの力学的変化（切り替わり方）を CONDENSATION に基づいてモデル化している[59]．一方，Dornaika らは，単一のパーティクルフィルタを用いて顔の追跡と表情認識を同時に行っている[60]．

線形動的システムの学習法としては，制御と同様にシステム同定を用いる方法[12),58] や，HMM と同様に EM アルゴリズムを用いる方法[61),62]，学習するパラメタを状態とみなしてカルマンフィルタによって推定する方法などがある[56]．

3.2.2 リカレントニューラルネット

リカレントニューラルネットは，非線形力学系として時系列パターン認識に用いられることがあり[63)~68]，

☆ 力学系と動的システムは共に dynamical system であるが，制御理論などで特に線形な系を対象とする場合は動的システムが，カオスなどの非線形の場合は力学系という言葉が用いられることが多いようである．なお，それぞれの論文の使用法に合わせ，ここでは具体的なモデルのことをシステムと呼ぶ場合がある．

脳のモデル化という観点からも，力学系モデルの上に離散事象系を構築することを目的として多くの興味深い研究が行われている．ただし，実際の応用には多くのハイパーパラメタを設定しておく必要がある．そのため，モデルの見通しのよさに欠け，ビジョンの分野（特に認識精度を競う研究）における事例は比較的少数に留まっている．

3.3 ハイブリッドシステム〜力学・離散事象混在系

力学系と離散事象系が混在した系をハイブリッドシステムと呼び，そのモデル化や解析方法について，1990 年前後から計算機科学や制御などの分野で盛んに研究行われるようになってきた[69)~71)]．ビジョンの分野ではこれとは独立に，人の構造的な行為や運動などを記述するためのモデルとして，力学系と離散事象系を組み合わせたシステムがいくつか提案されている．例えば，Siskind らは，缶などの物体を操作するときの力学的特徴を推論に組み込むことで，動的なシーンの解釈を与える方法などについて検討を行っている[72),73)]．

線形動的システムと HMM を直接組み合わせる方法として，カルマンフィルタの出力（例えば分散）を，HMM への入力とする方法もある[74)]．ただし以下では，両者がより密接に結びついたシステムに注目し，複数の線形動的システム（カルマンフィルタや自己回帰モデル，パーティクルフィルタ等）の間の遷移を HMM によってモデル化したものを取り上げる．

3.3.1 Switching linear dynamical system

線形動的システムと HMM を組み合わせたものとして，特に Bregler[55)] のモデルがよく知られており，画像上における人の足の動きを，複数の線形動的システムの切り替わりとして表現することで，skipping や hopping の認識などを行っている．つまり，複雑な足の動きを，単純な動き（足を前に出すなど）のフェーズの組み合わせとして考え，HMM の各状態が，それぞれ異なるフェーズ（この場合は 2 次の線形動的システム）に対応付けられている．

Dynamic Bayesian network の観点から両者を統合したものとして，HMM の状態に応じて状態空間自体を切り替える[75)] や，状態空間は共有する switching linear dynamical system (SLDS)[76),77)] があり，Murphy の文献[78)] の中でこれらの違いについて述べられている．

パーティクルフィルタと HMM を組み合わせたものとしては，前節で述べた Blake らの方法[59)] がある．

複数の HMM を用いて coupled HMM[41)] を構成したのと同様に，複数の SLDS において状態の共起関係をモデル化したものとして，coupled SLDS[79)] がある．また，HMM を parametric HMM[23)] へ拡張したのと同様に，SLDS に対して大域的な変動を表現するようなパラメタを導入したものとして，parametric SLDS[80)] がある．

3.3.2 持続時間を表現するハイブリッドシステム

前節で述べた，SLDS を始めとするモデルでは，観測サンプルが得られるごとに，HMM の離散的な状態と線形動的システムの連続的な状態の遷移同時に考える．すると，3.1.2 節で述べたように，1 つの離散状態が持続する長さは幾何分布に従うことになり，離散事象系にとっては不自然なモデルとなっている．そこで，離散事象系をサンプリング時刻に同期した遷移から切り離し，状態の持続長を明示的に分布としてモデル化したものとして，segment model[18)] や文献[81)] がある．

3.3.3 運動生成への応用

Bregler のモデルを始めとするこれらのシステムは生成モデルであり，時系列パターンの認識だけでなく，生成にも用いられることがある．あらかじめモーションキャプチャデータによって学習されたモデルから，人の複雑なダンスなどの動きを生成できるようにしたシステムとしては Motion texture[52)] がよく知られている．

3.3.4 ハイブリッドシステムの学習法

線形動的システムと HMM を組み合わせたハイブリッドシステムの学習では，HMM の状態数（線形動的システムの個数に一致）および，それぞれの線形動的システムのパラメタを推定する必要がある．歩行などの足や手の周期的動きの認識では，離散事象系の状態数が既知とする場合が多い．この場合，EM アルゴリズムによって反復的な学習を行うことによって，各線形動的システムのパラメタと，その間の遷移確率が推定される．また，行列演算を用いた学習も検討されている[82)]．

一方，状態数が未知とする場合は，いったん一定の範囲から推定した線形動的システムに，続く時系列データを入力（フィッティング）していき，フィッティング精度（尤度もしくはその近似として予測誤差）が閾値を超える場合に新たな線形動的システムを追加するといった，欲張り法によるアルゴリズムがある[52),83)]．ただし，ノイズが多いデータに対してはこの閾値の決め方が困難である．そこで，観測データ系列からボトムアップに線形動的システムの集合やその個数を学習する方法が提案されている[81),84)]．これは，複数の線形動的システムのうち性質の最も近いペアを順に併合する階層的クラスタリングであり，各システムの固有値が複素平面の単位円内に収まるように制約をかけることで，収束性の線形動的システムを，少ないデータから安定して学習可能である．状態数（線形動的システムの数）の決定は，通常の階層型クラスタリングと同様に，併合時の尤度（フィッティング精度）のカーブが急激に変化する時点をその候補としている．

4. 応 用 分 野

4.1 表 情 認 識

表情認識に関する研究は非常に多く，特にオプティカルフローに基づく動的な特徴量を抽出して，テンプレートマッチングや PCA，判別分析，SVM などの（静的）パターン認識手法を適用する方法がしばしば用いられる[85)~90)]．一方で，HMM などの時系列パターン認識手法を用いる手法も多く提案されている[91)~94)]．Cohen らは，静的なパターン認識と動的なパターン認識 (HMM) の比較を行っており，特定人物の表情を認識する場合や，表情を識別するための適当な時間区間の分節化が十分ではない場合には，静的なパターン認識手法に対して HMM が優位であるとコメントしている[95)]．また最近の研究では，3.2 節で述べたパーティクルフィルタを用いた表情認識が提案されている[60)]．

一方で，happy や sad といった感情的なカテゴリーではなく，社交的な笑いと自然な笑いといった微細な表情の認識では，人間は顔のパーツ（目や口など）の動きのタイミングやその持続時間を用いているという報告がある[96)]．そこで，ハイブリッドシステムによってあらかじめ各パーツの動きを分節化しておき，その動きの間のタイミングによって，自発的と意図的な笑いの判別を検討した研究がある[97)]．また，表情というよりは顔の動きから人の状態（覚醒，眠りなど）を認識するために，dynamic Bayesian network などが用いられている[98)]．

4.2 音響情報と視覚情報の統合

音声と動画像を統合することで，雑音環境下でも頑健な音声認識を実現させる audio-visual speech recognition(AVSR) の研究が，音声やビジョンなどの研究者によって行われている．特に OpenCV などでも用いられている，coupled HMM による方法[99)] や，dynamic Bayesian network を利用した手法[100)] などがよく知られている．これらは音声と動画像の間の共起関係を，隣り合う時刻の離散状態間の同時確率によってモデル化している．一方，音声から唇や顔の動画像を生成する研究もあり[101)]，音声で学習された HMM を用いて動画像の特徴ベクトル系列を学習することで，各状態において音声と動画像の対応をモデル化している．

参 考 文 献

1) Darrell, T. and Pentland, A.: Space-Time Gestures, *Proc. IJCAI'93 Looking at People Workshop* (1993).
2) Niyogi, S. and Adelson, E.: Analyzing and Recognizing Walking Figures in XYT, *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp.469–474 (1994).
3) Campbell, L. and Bobick, A.: Recognition of human body motion using phase space constraints, *Proc. IEEE Int. Conference on Computer Vision*, pp.624–630 (1995).
4) Ohno, H. and Yamamoto, M.: Gesture Recognition Using Character Recognition Techniques on Two-Dimensional Eigenspace, *Proc. IEEE Int. Conference on Computer Vision*, pp.151–156 (1999).
5) Chu, S., Keogh, E., Hart, D. and Pazzani, M.: Iterative Deepening Dynamic Time Warping for Time Series, *The Second SIAM Int. Conference on Data Mining* (2002).
6) Rabiner, L.R.: A tutorial on hidden Markov models and selected applications inspeech recognition, *Proc. IEEE*, pp.257–286 (1989).
7) Huang, H.D., Ariki, Y. and Jack, M.A.: *Hidden Markov Models for Speech Recognition*, Edinburgh Univ. (1990).
8) 中川聖一：音声認識研究の動向，電子情報通信学会論文誌， Vol. J83-D-II, No. 2, pp. 433–457 (2000).
9) Bobick, A.F. and Wilson, A.D.: A State-Based Approach to the Representation and Recognition of Gesture, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 19, No. 12, pp. 1325–1337 (1997).
10) Yamato, J., Ohya, J. and Ishii, K.: Recognizing Human Action in Time-Sequential Images Using Hidden Markov Model, *Proc. Int. Conference on Computer Vision and Pattern Recognition*, pp.379–385 (1992).
11) Moore, D. and Essa, I.: Recognizing multitasked activities from video using stochastic context-free grammar, *Proc. National Conference on Artificial intelligence*, pp. 770–776 (2002).
12) Doretto, G., Chiuso, A., Wu, Y.N. and Soatto, S.: Dynamic Textures, *Int. Journal of Computer Vision*, Vol.51, No.2, pp.91–109 (2003).
13) Wilson, A. D., Bobick, A. F. and Cassell, J.: Temporal Classification of Natural Gesture with Application to Video Coding, *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp.948–954 (1997).
14) Levinson, S. E.: Continuously variable duration hidden Markov models for automatic speech recognition, *Computer Speech and Language*, Vol.1, pp.29–45 (1986).
15) Wada, T. and Matsuyama, T.: Multiobject Behavior Recognition by Event Driven Selective Attention Method, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol.22, No.8, pp.873–887 (2000).
16) 大和淳司，大谷　淳，石井健一郎：隠れマルコ

フモデルを用いた動画像からの人物の行動認識, 電子情報通信学会論文誌, Vol.J76-D-II, No.12, pp.2556–2563 (1993).
17) Starner, T. and Pentland, A.: Visual Recognition of American Sign Language Using Hidden Markov Models, *Proc. Int. Workshop on Automatic Face and Gesture Recognition*, pp. 189–194 (1995).
18) Ostendorf, M., Digalakis, V. and Kimball, O.A.: From HMMs to Segment Models: A Unified View of Stochastic Modeling for Speech Recognition, *IEEE Trans. Speech and Audio Process*, Vol.4, No.5, pp.360–378 (1996).
19) Marcel, S., Bernier, O., Viallet, J.-E. and Collobert, D.: Hand Gesture Recognition Using Input-Output Hidden Markov Models, *Proc. IEEE Int. Conference on Automatic Face and Gesture Recognition*, pp.456–461 (2000).
20) Just, A., Bernier, O. and Marcel, S.: HMM and IOHMM for the Recognition of Mono- and Bi-Manual 3D Hand Gestures, *British Machine Vision Conference* (2004).
21) Wilson, A. and Bobick, A.: Learning Visual Behavior for Gesture Analysis, *IEEE International Symposium on Computer Vision*, p. 5A Motion II (1995).
22) Hamdan, R., Heitz, F. and Thoraval, L.: Gesture Localization and Recognition using Probabilistic Visual Learning, *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp.98–103 (1999).
23) Wilson, A. D. and Bobick, A. F.: Nonlinear PHMMs for the Interpretation of Parameterized Gesture, *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp.879–884 (1998).
24) Wilson, A. D. and Bobick, A. F.: Parametric hidden Markov models for gesture recognition, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol.21, No.9, pp.884–900 (1999).
25) Elgammal, A., Shet, V., Yacoob, Y. and Davis, L.S.: Learning Dynamics for Exemplar-based Gesture Recognition, *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp.571–578 (2003).
26) Deligne, S. and Bimbot, F.: Inference of Variable-Length Linguistic and Acoustic Units by Multigrams, *Speech Communication*, Vol.23, pp.223–241 (1997).
27) Murphy, K. P.: Hidden semi-Markov models (HSMMs), *Informal Notes* (2002).
28) Hongeng, S. and Nevatia, R.: Large-Scale Event Detection Using Semi-Hidden Markov Models, *Proc. IEEE Int. Conference on Computer Vision*, pp.1455–1462 (2003).
29) Duong, T. V., Bui, H. H., Phung, D. Q. and Venkatesh, S.: Activity Recognition and Abnormality Detection with the Switching Hidden Semi-Markov Model, *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp.838–845 (2005).
30) Brants, T.: Estimating HMM topologies, *Logic and Computation* (1995).
31) Brants, T.: Better Language Models with Model Merging, *Proc. the Conference on Empirical Methods in Natural Language Processing* (1996).
32) Brants, T.: Estimating Markov Model Structure, *Proc. Int. Conference on Spoken Language Processing*, pp.893–896 (1996).
33) Gong, S., Walter, M. and Psarrou, A.: Recognition of Temporal Structures: Learning Prior and Propagating Observation Augmented Densities via Hidden Markov States, *Proc. IEEE Int. Conference on Computer Vision*, pp.157–162 (1999).
34) Brand, M. and Kettnaker, V. M.: Discovery and Segmentation of Activities in Video, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol.22, No.8, pp.844–851 (2000).
35) Fine, S., Singer, Y. and Tishby, N.: The Hierarchical Hidden Markov Model: Analysis and Applications, *Machine Learning*, Vol.32, No.1, pp.41–62 (1998).
36) Bui, H.H., Phung, D.Q. and Venkatesh, S.: Hierarchical Hidden Markov Models with General State Hierarchy, *Proc. National Conference on Artificial Intelligence*, pp.324–329 (2004).
37) Nguyen, N. T., Phung, D. Q., Venkatesh, S. and Bui, H.: Learning and Detecting Activities from Movement Trajectories Using the Hierarchical Hidden Markov Model, *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp.955–960 (2005).
38) Oliver, N., Garg, A. and Horvitz, E.: Layered representations for learning and inferring office activity from multiple sensory channels, *Computer Vision and Image Understanding*, Vol.96, No.2, pp.163–180 (2004).
39) Ron, D., Singer, Y. and Tishby, N.: The power of amnesia: Learning Probabilistic Automata with Variable Memory Length, *Machine Learning*, Vol.25, No.2-3, pp.117–149 (1996).
40) Galata, A., Johnson, N. and Hogg, D.: Learning Variable-Length Markov Models of Behavior, *Computer Vision and Image Understanding*, Vol.81, No.3, pp.398–413 (2001).
41) Brand, M., Oliver, N. and Pentland, A.: Cou-

pled Hidden Markov Models for complex action recognition, *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp.994–999 (1997).
42) Juang, B.H. and Rabiner, L.R.: A probabilistic distance measure for hidden Markov models, *AT & T Technical Journal*, Vol.64, No.2, pp.391–408 (1985).
43) Alon, J., Sclaroff, S., Kollios, G. and Pavlovic, V.: Discovering clusters in motion time-series data, *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 375–381 (2003).
44) Ghahramani, Z.: Learning Dynamic Bayesian Networks, *Lecture Notes in Artificial Intelligence. Springer-Verlag.*, Vol.1387, pp.168–197 (1998).
45) Ghahramani, Z. and Jordan, M. I.: Factorial Hidden Markov Models, *Machine Learning*, Vol.29, No.2-3, pp.245–273 (1997).
46) Murphy, K.P.: Dynamic Bayesian Networks: Representation, Inference and Learning, *PhD Thesis, UC Berkeley, Computer Science Division* (2002).
47) Ayers, D. and Chellappa, R.: Scenario recognition from video using a hierarchy of dynamic belief networks, *Proc. Int. Conference on Pattern Recognition*, pp.835–838 (2000).
48) Mittal, A., Cheong, L.F. and Sing, L.T.: Dynamic Bayesian framework for extracting temporal structure in video, *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp.110–115 (2001).
49) Xiang, T. and Gong, S.: Beyond Tracking: Modelling Activity and Understanding Behaviour, *International Journal of Computer Vision*, Vol.67, No.1, pp.21–51 (2006).
50) Ivanov, Y.A. and Bobick, A.F.: Recognition of visual activities and interactions by stochastic parsing, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol.22, No.8, pp.852–872 (2000).
51) Minka, T. P.: From Hidden Markov Models to Linear Dynamical Systems, *MIT Media Lab Vision and Modeling TR531* (1999).
52) Li, Y., Wang, T. and Shum, H.-Y.: Motion Texture: A Two-Level Statistical Model for Character Motion Synthesis, *Proc. SIGGRAPH*, pp.465–472 (2002).
53) Anderson, B. D.O. and Moor, J.B.: *Optimal Filtering*, Prentice-Hall (1979).
54) Isard, M. and Blake, A.: Condensation – conditional density propagation for visual tracking, *Int. Journal of Computer Vision*, Vol.29, No.1, pp.5–28 (1998).
55) Bregler, C.: Learning and Recognizing Human Dynamics in Video Sequences, *Proc. Int. Conference on Computer Vision and Pattern Recognition*, pp.568–574 (1997).
56) Rao, R.: Dynamic Appearance-Based Recognition, *Proc. Int. Conference on Computer Vision and Pattern Recognition*, pp. 540–546 (1997).
57) Rittscher, J. and Blake, A.: Classification of Human Body Motion, *Proc. IEEE Int. Conference on Computer Vision*, pp.634–639 (1999).
58) Bissacco, A., Chiuso, A., Ma, Y. and Soatto, S.: Recognition of Human Gaits, *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp.52–57 (2001).
59) Blake, B. N.A., Isard, M. and Rittscher, J.: Learning and Classification of Complex Dynamics, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol.22, No.9, pp.1016–1034 (2000).
60) Dornaika, F. and Davoine, F.: Simultaneous Facial Action Tracking and Expression Recognition Using a Particle Filter, *Proc. IEEE Int. Conference on Computer Vision*, pp.1733–1738 (2005).
61) Ghahramani, Z. and Hinton, G.E.: Parameter Estimation for Linear Dynamical Systems, *Technical Report CRG-TR-96-2* (1996).
62) Roweis, S. and Ghahramani, Z.: A Unifying Review of Linear Gaussian Models, *Neural Computation*, Vol.11, No.2, pp.305–345 (1999).
63) 二見亮弘，星宮　望：時系列パターン認識の神経回路モデル，電子情報通信学会論文誌， Vol.J71-D-II, No.10, pp.2181–2190 (1988).
64) Elman, J.L.: Finding Structure in Time, *Cognitive Science*, Vol.14, pp.179–211 (1990).
65) Robinson, A.J.: An Application of Recurrent Nets to Phone Probability Estimation, *IEEE Trans. Neural Networks*, Vol.5, No.2, pp.298–305 (1994).
66) Dorffner, G.: Neural Networks for Time Series Processing, *Neural Network World*, Vol.6, No.4, pp.447–468 (1996).
67) 森田昌彦，村上　聡：非単調神経回路網による時系列パターンの認識，電子情報通信学会論文誌，Vol.J81-D-II, No.7, pp.1679–1688 (1998).
68) 内山　徹，高橋治久：リカレントニューラル予測モデルを用いた不特定話者単語音声認識，電子情報通信学会論文誌， Vol.J83-D-II, No.2, pp.776–783 (2000).
69) Gollu, A. and Varaiya, P.: Hybrid Dynamical Systems, *Proc. Conference on Decision and Control*, pp.2708–2712 (1989).

70) Maler, O., Manna, Z., de Bakker, A. P. J.W., Huizing, C., de Roever, W. P. and Rozenberg), G.: *From Timed to Hybrid Systems*, Real-Time: Theory in Practice, pp. 447–484, Springer-Verlag (1991).
71) R.Alur, e.a.: The Algorithmic Analysis of Hybrid Systems, *Theoretical Computer Science*, Vol.138, No.1, pp.3–34 (1995).
72) Mann, R., Jepson, A. and Siskind, J.M.: The Computational Perception of Scene Dynamics, *Computer Vision and Image Understanding*, Vol.65, No.2, pp.113–128 (1997).
73) Siskind, J.M.: Reconstructing force-dynamic models from video sequences, *Artifitial Intelligence*, Vol.151, No.1-2, pp.91–154 (2002).
74) Dockstader, S.L., Imennov, N.S. and Tekalp, A.M.: Markov-Based Failure Prediction for Human Motion Analysis, *Proc. IEEE Int. Conference on Computer Vision*, pp. 1283–1288 (2003).
75) Ghahramani, Z. and Hinton, G.E.: Switching State-Space Models, *Dept. of Computer Science* (1996).
76) Pavlovic, V., Rehg, J.M., Cham, T. and Murphy, K.P.: A Dynamic Bayesian Network Approach to Figure Tracking Using Learned Dynamic Models, *Proc. IEEE Int. Conference on Computer Vision*, pp.94–101 (1999).
77) Pavlovic, V., Rehg, J.M. and MacCormick, J.: Impact of Dynamic Model Learning on Classification of Human Motion, *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp.788–795 (2000).
78) Murphy, K.P.: Switching Kalman Filter, *Technical report, U. C. Berkeley* (1998).
79) Jeong, M. H., Kuno, Y. and Shimada, N.: Two-Hand Gesture Recognition using Coupled Switching Linear Model, *Proc. Int. Conference on Pattern Recognition*, pp.529–532 (2002).
80) Oh, S. M., Rehg, J. M., Balch, T. and Dellaert, F.: Learning and Inference in Parametric Switching Linear Dynamical Systems, *Proc. IEEE Int. Conference on Computer Vision*, pp. 1161–1168 (2005).
81) Kawashima, H. and Matsuyama, T.: Multiphase Learning for an Interval-based Hybrid Dynamical System, *IEICE Trans. Fundamentals*, Vol.E88-A, No.11, pp.3022–3035 (2005).
82) Bissacco, A.: Modeling and Learning Contact Dynamics in Human Motion, *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp.421–428 (2005).
83) Lu, C., Liu, H. and Ferrier, N.J.: Multidimensional Motion Segmentation and Identification, *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp.2629–2636 (2000).
84) Kawashima, H. and Matsuyama, T.: Hierarchical Clustering of Dynamical Systems based on Eigenvalue Constraints, *3rd International Conference on Advances in Pattern Recognition (S. Singh et al. (Eds.): ICAPR 2005, LNCS 3686)*, pp.229–238 (2005).
85) Essa, I.A. and Pentland, A.P.: Facial Expression Recognition using a Dynamic Model and Motion Energy, *Proc. IEEE Int. Conference on Computer Vision*, pp.360–367 (1995).
86) Yacoob, Y. and Davis, L.S.: Recognizing Human Facial Expressions from Long Image Sequences Using Optical-Flow, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol.18, No.6, pp.636–642 (1996).
87) Essa, I.A. and Pentland, A.P.: Coding, Analysis, Interpretation, and Recognition of Facial Expressions, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol.19, No.7, pp. 757–763 (1997).
88) Kimura, S. and Yachida, M.: Facial Expression Recognition and Its Degree Estimation, *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp.295–300 (1997).
89) Donato, G., Bartlett, M.S., Hager, J.C., Ekman, P. and Sejnowski, T.: Classifying facial actions, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 21, No. 10, pp. 974–989 (1999).
90) Bartlett, M. S., Littlewort, G., Frank, M., Lainscsek, C., Fasel, I. and Movellan, J.: Recognizing Facial Expression: Machine Learning and Application to Spontaneous Behavior, *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp.568–573 (2005).
91) Otsuka, T. and Ohya, J.: Recognizing Abruptly Changing Facial Expressions from Time-Sequential Face Images, *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp.808–813 (1998).
92) Hoey, J. and Little, J.J.: Representation and Recognition of Complex Human Motion, *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp.752–759 (2000).
93) Chang, Y., Hu, C. and Turk, M.: Probabilistic Expression Analysis on Manifolds, *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp.520–527 (2004).
94) Yeasin, M., Bullot, B. and Sharma, R.: From Facial Expression to Level of Interest: A Spatio-Temporal Approach, *Proc. IEEE Conference on Computer Vision and Pattern*

Recognition, pp.922–927 (2004).
95) Cohen, I., Sebe, N., Garg, A., Chen, L.S. and Huang, T.S.: Facial expression recognition from video sequences: temporal and static modeling, *Computer Vision and Image Understanding*, Vol.91, No.1-2, pp.160–187 (2003).
96) Nishio, S., Koyama, K. and Nakamura, T.: Temporal Differences in Eye and Mouth Movements Classifying Facial Expressions of Smiles, *Proc. IEEE Int. Conference on Automatic Fase and Gesture Recognition*, pp.206–211 (1998).
97) Nishiyama, M., Kawashima, H., Hirayama, T. and Matsuyama, T.: Facial Expression Representation based on Timing Structures in Faces, *IEEE International Workshop on Analysis and Modeling of Faces and Gestures (W. Zhao et al. (Eds.): AMFG 2005, LNCS 3723)*, pp.140–154 (2005).
98) Gu, H. and Ji, Q.: Facial Event Classification with Task Oriented Dynamic Bayesian Network, *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 870–875 (2004).
99) Nefian, A.V., Liang, L., Pi, X., Xiaoxiang, L., Mao, C. and Murphy, K.: A Coupled HMM for Audio-Visual Speech Recognition, *Proc. IEEE Int. Conference on Acoustics, Speech and Signal Processing*, Vol.2, pp.2013–2016 (2002).
100) Nefian, A. V., Liang, L., Pi, X., Liu, X. and Murphy, K.: Dynamic Bayesian Networks for Audio-Visual Speech Recognition, *EURASIP Journal on Applied Signal Processing*, Vol.2002, No.11, pp.1–15 (2002).
101) Brand, M.: Voice Puppetry, *Proc. SIGGRAPH*, pp.21–28 (1999).