社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report
TL2006-71,SP2006-159,WIT2006-103(2007-01)

状態遷移構造の推定に基づく手話認識

松尾　直志†　　白井　良明†

† 立命館大学情報理工学部　〒525–8577 滋賀県草津市野路東 1–1–1
E-mail: †matsuo@i.ci.ritsumei.ac.jp, ††shirai@ci.ritsumei.ac.jp

あらまし　本報告では手話単語動作を撮影した画像列から単語を認識する際に HMM を用いる手法について考える．HMM を用いる認識手法は既にいくつか提案されているが，従来法ではモデルの状態遷移構造は単語に関係なく固定されていた．しかし手話単語動作をいくつかの動きに分割したとき，同じ意味をもつ動作でもその中の動きは省略されていたり変形している場合がある．発話ごとに変わる動きを認識するには起こり得る動きの系列を求めそれに応じた状態遷移構造を用いなければならない．そこで本報告では複数の学習用画像列から，起こり得る動き系列を求め状態遷移構造を推定する方法について提案する．実験の結果，手話動作のパターンに応じた状態遷移構造が得られることが確認できた．

キーワード　手話認識，動画像処理，HMM，動き抽出，状態遷移推定

Sign Language Recognition based on Estimation of Transition Structure

Tadashi MATSUO† and Yoshiaki SHIRAI†

† College of Information Science and Engineering, Ritsumeikan University　Noji-higashi 1–1–1, Kusatsu-shi, Shiga, 525–8577 Japan
E-mail: †matsuo@i.ci.ritsumei.ac.jp, ††shirai@ci.ritsumei.ac.jp

Abstract　In this report, we describe a method using HMM to recognize sign language from an image sequence. Some methods using HMM have been already proposed, but the topology of their HMMs is fixed for all signs. We should use topology suitable for dynamics of each sign because some motions in a sign may be skipped or modified. Accordingly, we propose a method to extract and match motions from multiple image sequences for learning and a method to estimate suitable topology. By some experiments we show that the proposed method works well.

Key words　sign recognition, video processing, HMM, motion extraction, estimation of transitional topology

1. はじめに

手話の一連の動作の中には「手を上げる」,「手を広げる」などの複数フレームにまたがるまとまった動きがあるため，この各動きをひとつの状態に対応させた隠れマルコフモデル（Hidden Markov Model, HMM）[1] による認識が有効と考えられる．これまでにも画像から抽出した特徴量を基にした HMM による手話認識法が研究されている [2]〜[4]．しかし従来研究では HMM の構造が単語によらず固定されており，またある動きに対しては有効な特徴量が，他の動きに対しては不適切になるという性質を考慮していなかった．本報告ではこれらの点を考慮した HMM を自動的に構成する方法について述べる．

2. 従来法の問題点

HMM を学習する際には前もってその状態数と遷移構造を決定する必要がある．HMM の構造については実験から経験的に求めるという方法 [2] もあるが，手話単語ごとに動作の複雑さや各状態間に許される遷移関係は異なるので単語の動作に応じた構造を用いるのが望ましい．学習用データから状態数を求める方法としては，手領域の移動速度を主に用いる方法 [3] が提案されているが，速度がほとんど変化しない場合の状態遷移を抽出できず，また話者によって手話動作の速度が異なるため，話者ごとに閾値が必要という問題点があった．また，状態間の遷移構造としては図 1 のような単純な left-to-right 構造 [3] や，省略を許した図 2 のような構造 [2] が提案されている．しかし，従来これらの構造自体は単語によらず固定されており，実際の手話単語に現れる繰り返し動作や，一部動作の省略等に対応できないものとなっていた．

また，HMM では各状態に属しているときにはある一定の分布をもつ出力が得られることが想定されているが，動きについての特徴量は一般にはそうはならない．図 3 は「手を上げる」動作を行っているときの，手重心の Y 座標の変化を示したもの

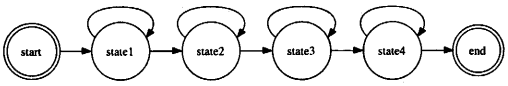


図 1　単純な状態遷移構造

Fig. 1　A simple transitional topology.

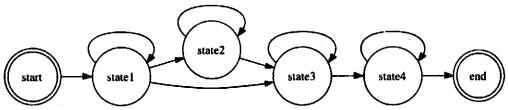


図 2　省略を許容できる遷移構造

Fig. 2　A topology with a skip.

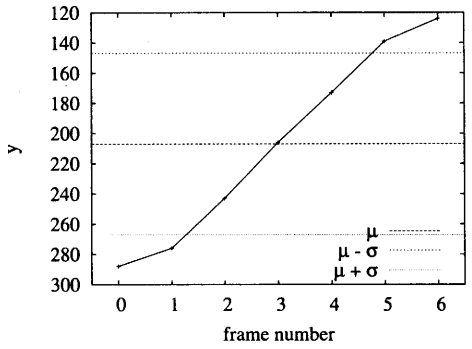


図 3　「手を上げる」動作時の手重心 Y 座標

Fig. 3　Y coordinate of hand in motion.

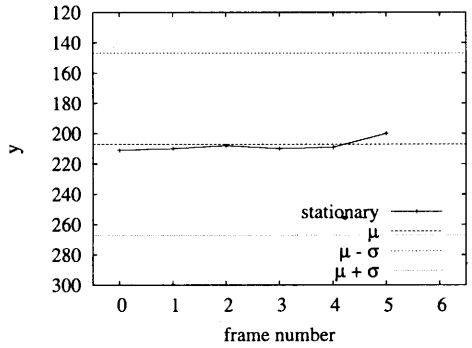


図 4　図 3 の統計量と静止時の手重心 Y 座標の関係

Fig. 4　Y coordinate of hand in stationary.

である．図中の μ, σ はそれぞれこの区間での Y 座標値の平均と標準偏差を表している．この図からも明らかなように，「手を上げる」という動作においては，「各フレームでの Y 座標値が平均に近い」ということが「手を上げる動作の途中である」ということに直接つながっていない．実際，この動きの場合には「手重心位置が上に動いている」状態であるので「各フレームの位置」は重要でないばかりか，図 3 での Y 座標値がガウス分布に従うものとして平均と分散から「手を上げている状態」の尤度を求めると，「手を静止させている」場合の方が「手を上げている」尤度が高いという現象が起こる．この状況を示したのが図 4 であう．図 4 中の μ, σ は図 3 の「手を上げる」動作時の Y 座標の平均と分散である．図 3 の「手を上げる」動作の場合よりも平均に近い値となっているので尤度が高いと見なされてしまう．ある区間内での重心位置平均と分散から求めた，つまりガウス分布を仮定した尤度が本来の動作の類似度に対応していないことになる．

従来法の HMM の学習では，初期モデルが与えられたときにその各状態に学習用データの各フレームを割り振って状態のパラメータを更新していた．各状態へのフレームの割り振りは尤度を最大とするように行われるため，尤度が動作の類似度に対応していないと正しく学習を行うことができない．上記の例では，重心位置を特徴量に含めると「手を上げる」状態の学習に「手を静止させている」フレームの特徴量を用いてしまう可能性がある．

この場合には位置の値そのものを特徴量としなければよいが，一方で指で顔の部位を指し示す場合など，位置が重要となる手話も存在する．従来法のように固定した構造をもつ HMM をもとにした単純な学習では，状態ごとに用いるべき特徴量が異なる状況に対応できない．位置情報を特徴量に含めたとしても，移動状態では重心位置が尤度にほとんど寄与しないようにパラメータを調整できれば上記のような問題は回避できると思われる．そのような調整を行うためには複数の画像列から似た動きを行っている部分を抽出し，それらの情報を総合して動きの属性に応じた処理を施す必要がある．手話の中には静止状態が入る場合と入らない場合のどちらもが起こり得る単語や，発話によって動作の一部が変形する動きがあるため，ある画像列には含まれている動きが別の画像列には含まれていないこともある．このため，似た動きの対応付けを行う際にその手話単語で許される状態遷移をも考慮しなければならない．以下，3. でまず学習用データをひとまとまりの動きに対応した区間に分割する方法について述べ，続く 4. で複数の区間列から対応関係を求めて状態遷移構造を推定する方法について述べる．

3.　学習用データの区間分割

状態遷移構造を推定するため，学習用データを「手を上げる」等のひとまとまりの動きを単位とした区間に分割する．ここではまず，片手についての手領域重心の速度と動きの向きを基にした以下のような手法で意味のある動きの区間を抽出する．

（1）　重心位置の移動速度が閾値 v_0 以下の場合は静止区間とする．静止区間以外の区間を移動区間とする．

（2）　各移動区間内で重心位置の動きの向きの角度を求め，同一区間内の速度最大フレームとの角度差を計算する．

（3）　動きの向きの角度差が閾値 θ_0 以上となるフレームが同じ区間内にあればその移動区間を分割する．動きの向きの差が θ_0 以下になるまで分割を繰り返す．

（4）　属するフレーム数が閾値以下，あるいは区間内での総移動距離が閾値以下の区間は過渡区間とし，連続する過渡区間はひとつにまとめる．

この手順によって学習用データは「手が静止している」静止区間，「手がある方向に動いている」移動区間，「手は動いているものの向きの変化が激しい」過渡区間の 3 つに分類される．

左右の手を組み合わせて状態を構成するため，それぞれの手の動きについて求めた区間列の同期を取る必要がある．ここでは図 5 のように，左右の区間が十分長くフレームを共有する場

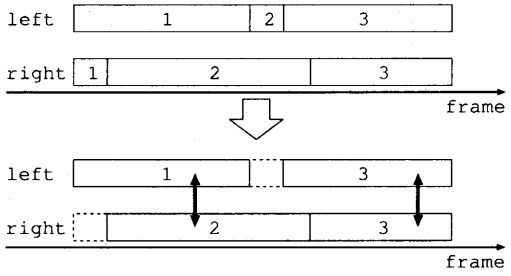


図 5　左右の手についての区間の同期
Fig. 5　Synchronization of intervals of hands.

合，あるいは左右の区間が共有するフレーム内での総移動距離が十分長い区間をペアとする．ペアのつくれない区間はフレーム数も少なく，移動距離も短いうえ，動作のタイミングも揃っていないので重要でないと見なして除外する．このようにしてひとまとまりの動きについての情報を求めることができる．

4.　状態遷移構造の推定

これまでに述べた手法で，ひとつの学習用画像列からひとまとまりの動きの区間列を得ることができる．この区間列のそれぞれを状態に対応させれば直接に HMM を生成できるが，同一手話単語で異なる画像列から得た区間列についても考慮する必要がある．ここではある区間列から初期モデルを構成し，残りの区間列の情報を初期モデルに追加していくという方法を用いる．異なる区間列からの情報を合成するには，モデル中の状態がどの区間の動きと対応するかを求めればよい．例として図 6 上段のようなモデルが得られているときに図 6 中段で示される区間列を合成することを考える．このとき，区間列は静止区間，移動区間，静止区間が並んでいるが，モデルは移動状態と静止状態の 2 つのみであるため，最初の静止区間はモデル中に対応する状態が存在しない．このことから，このモデルには最初の静止区間に相当する状態が含まれていないといえるので，最初の静止区間に相当する状態を新しい経路として追加すれば，この区間列と対応させることができる（図 6 下段）．

このように，モデル中の状態と与えられた区間列のそれぞれとの対応付けができれば，与えられた区間列を許容するようモデルを更新していくことができる．対応付けの手法としては，全ての状態と区間の間にスコアを定義し，スコアの総合計が最大となるような状態，区間のペアの系列を求めてそれを採用した．但し，区間列としては与えられた順序を保つものを，状態列としてはモデルの遷移関係を見たすもののみについてスコアを考えている．スコアの値としては以下のような値を用いた．

$$\mathrm{score}(state, segment) = \begin{cases} 1.0, & \text{（静止状態と静止区間）} \\ -1.0, & \text{（静止状態と移動区間，あるいは移動状態と静止区間）} \\ \cos(\theta_{segment} - \theta_{state}), & \text{（移動状態と移動区間，移動状態と過渡区間）} \end{cases} \quad (1)$$

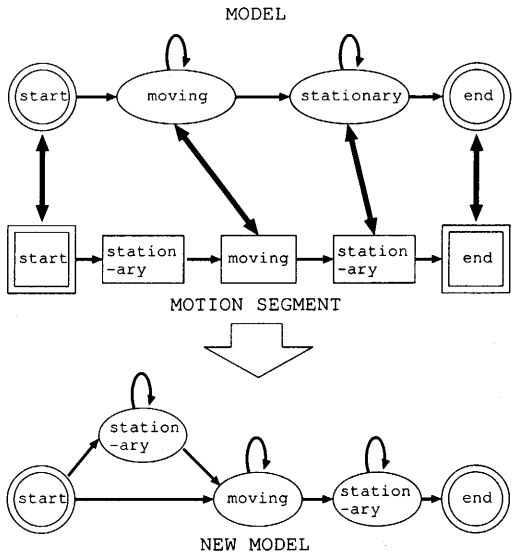
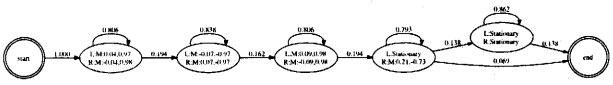


図 6　区間列との照合を用いた状態の追加
Fig. 6　Addition of state by matching with segments.

(a) 終了前に静止状態なし



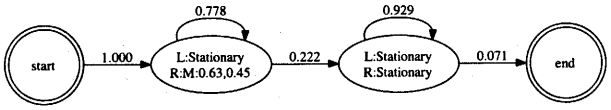
(b) 終了前に静止状態をもつ区間列を追加

図 7　状態追加の例（「暖かい」）
Fig. 7　An example of addition of state (“warm”).

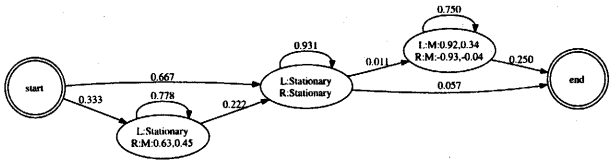
ここで $\theta_{state}, \theta_{segment}$ は，注目する状態や区間での，動きの向きの平均である．移動区間としてはほぼ一定の向きに移動する区間を抽出しているので，2. での位置情報とは異なり平均が意味をもつ．このようなスコアを用いることで，静止区間は静止状態と，移動区間は向きの動きの向きの近い移動状態と対応させることができる．

実際の手話画像から手重心位置の動きを抽出し，前述の方法で状態遷移構造を推定した結果が図 7 である．この手話は「暖かい」を表す手話で，両手を胸の前まで上げた後，両手をほぼ同時に上下させる動作である．実験に用いた画像列では，手を上下させる動作の後にすぐ終了し静止状態のないものと，動作の後に静止状態が入るものとがある．図 7(a) は静止状態のない画像列から得られた構造であるが，これに静止状態をもつ画像列の情報を合成することで図 7(b) のように静止状態が入る場合も許容できるように状態が追加される．

また，「合う」という意味の手話単語について推定を行った結果が図 8 である．「合う」という単語の場合，左右の手の指先を



(a) 指を合わせて静止する



(b) 指を合わせた後に少し動く区間列を追加

図 8　状態追加の例（「合う」）
Fig. 8　An example of addition of state("match").

上下から合わせるという動作になっている．図 8(a) は指先を合わせる前の動きと，合わせた後の静止状態が検出された画像列から生成した状態遷移構造である．同図 (b) は合わせる前の動作が小さいため検出できていないような画像列や，指先を合わせた状態で少し動くのが検出される画像列の情報を合成したものである．

このようにして状態遷移構造を構成すると各状態に属するフレームも同時に求まるので，各状態がもつ特徴量についてのパラメータ（平均，分散）も容易に求められる．但し，2. で指摘したように状態によっては平均と分散を基にして尤度を計算する，つまりガウス分布を仮定するのが適切でない場合もある．不適切な特徴量については分散を非常に大きく設定し，その特徴量の大小が尤度に与える影響を弱める手法が考えられるが，どの特徴量についてこの処理を行うか，行うならばどの程度弱めるべきかについて検討する必要がある．

5. 認識実験

提案手法を実際の手話画像列に適用し，認識実験を行った．簡単のため特徴量としては動きの向きを表す単位ベクトルと，速度の逆数を用いた．提案法との比較のため，遷移構造は図 1 のような単純な left-to-right 構造に限定したうえで単一の学習データから状態数のみを推定してモデルを生成した場合についての実験も行った．この単純な left-to-right 構造の場合には，図 9 のような軌跡を持つ「合う」という手話単語が図 10 のような軌跡を持つ「セーター」という手話単語と誤認識された．

図 9，10 の軌跡では明らかに異なる動きであるが誤認識されてしまうのは，同じ単語を意味する手話であっても，各画像列のもつ動きが異なるために学習が適切なものになっていないためと考えられる．「セーター」という単語の場合，手を胸の前から下げる動作があり，下げる前に静止状態が入る画像列も入らない画像列も起こり得る．このため，抽出できる区間列としては図 11 のようにいくつかのパターンが現れる．図 11 中の四角形はひとつの区間を表しており，四角形の中の L と R がそれぞれ左手，右手を意味し，M と S がそれぞれ移動状態と静止状態に表している．仮に図 11(a) の区間列から初期モデルを生成し，単純な left-to-right 構造を仮定すると状態遷移構造は移動状態から静止状態へ，という構造に固定される．このように構造を固定して学習を行うと移動状態の学習には各画像列の先頭の数フレームを用いることになる．このため図 11(b) の区間列の最初に現れる静止区間に属する数フレームが移動状態のパラメータを学習するのに採用されてしまう．実際には静止区間であるので移動状態の学習には不適切であるが，対応する静止区間が見つからないため移動状態に割り振られてしまう．移動状態と静止状態では得られる特徴量が異なるので，モデル内の各状態と各フレームとの対応抽出（Viterbi アルゴリズム）の

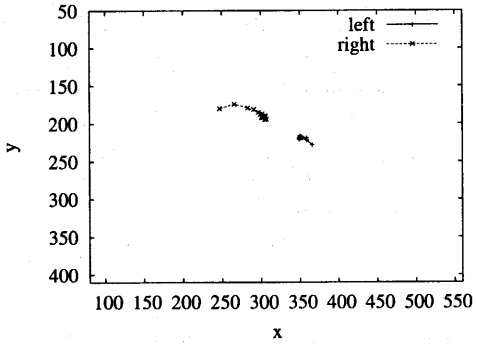


図 9　「合う」の軌跡
Fig. 9　A trajectory of hands in sign, "match".

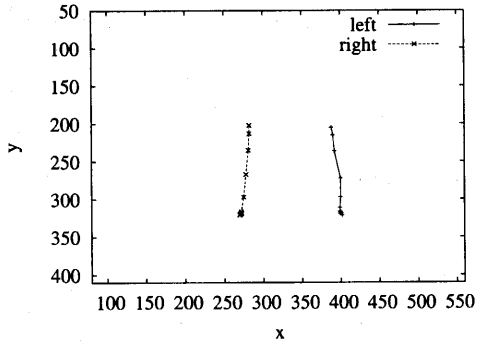
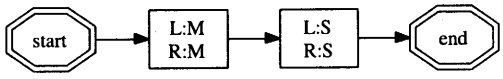


図 10　「セーター」の軌跡
Fig. 10　A trajectory of hands in sign, "sweater".



(a) 動作の前に静止状態はない場合

(b) 動作の前に静止状態がある場合

図 11　同一手話単語のバリエーション（「セーター」）
Fig. 11　Variations of sign with the same meaning ("sweater").

通信学会技術研究報告, 第 104 巻, pp. 21–28 (2004).

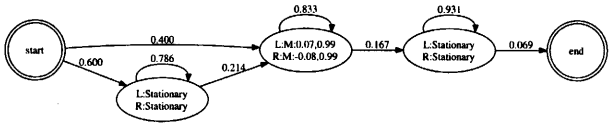


図 12 「セーター」の状態遷移構造
Fig. 12 The estimated topology for "sweater".

際に移動状態に静止フレームが割り振られないようにする効果は働くが，対応する状態がそもそも存在しなければ，適切な対応関係は得られない．このため初期モデルでは移動状態であった状態が移動状態と静止状態を混ぜ合わせたパラメータとなってしまい，手話動作を代表する状態ではなくなってしまうと考えられる．

一方，提案法の場合にはあらかじめ複数の学習用データから起こり得る動きの情報を総合してモデルを生成する (図 12) ので，学習時の状態の混乱は起こらない．上記の「合う」についても正しく認識することができた．また今回の実験では特別な調整は行っていないが，モデルのパラメータを定める際にも，移動状態であるか静止状態であるかといった動きの属性を用いて，不適切なパラメータ（移動状態における各フレームでの重心位置など）が尤度に悪影響を及ぼさないように調整することも可能である．

6. おわりに

本報告では HMM を用いた手話認識における従来法の問題について述べ，それを解決するため対応する動きを自動的に求める方法を提案した．提案法により同じ意味をもつ手話単語の動作の多様性に応じた状態遷移構造を推定することができる．実験の結果，動きの大きい手話単語については概ね正しい遷移構造が求められているようである．今回の実験では特徴量として動きの向きと速度を用いたが，単語によっては位置が重要であるため，位置に関する特徴量も導入する必要がある．その場合，移動状態においては位置特徴量が尤度に寄与しないようパラメータを調整しなければならない．その際に必要となる，調整すべき状態であるかどうかの判定法や，調整するとしてどのように行うべきか，どの程度寄与を弱めるべきかについて今後検討する予定である．また，手を震わせる動作や扇ぐ動作など，手重心位置の移動速度が閾値に近い動作が続く場合に，非常に短い静止状態と移動状態が繰り返し検出され状態数が増えすぎでしまうという問題もある．この場合には微小な動きの繰り返しが意味をもつので，一度分割した区間列を再度統合する必要がある．この点についても研究を行う予定である．

文　献

[1] 中川："確率モデルによる音声認識", 電子情報通信学会，コロナ社 (1988).
[2] T. Starner, J. Weaver and A. Pentland: "Real-time american sign language recognition using desk and wearable computer based video", IEEE Transactions on Pattern Analysis and Machine Intelligence, **20**, 12, pp. 1371–1375 (1998).
[3] 川東, 白井, 島田, 三浦："手話の HMM 作成のための状態分割", 電子情報通信学会技術研究報告, No. 67, pp. 55–60 (2005).
[4] 金山, 白井, 島田："HMM を用いた手話単語の認識", 電子情報