H-021

階層隠れ CRF によるスポーツ映像のセグメンテーション
Sports Video Segmentation by Hierarchical Hidden CRF

玉田 寛尚 †
Tamada Hirotaka

林 朗 †
Hayashi Akira

1 はじめに

階層隠れマルコフモデル (HHMM) は，時系列データのモデルとしてよく知られている HMM の状態空間に階層構造を持たせたモデルである．近年，生成モデルである HMM に対して，識別モデルである条件付確率場 (CRF) が提案され，有効性が確認されている．筆者らも生成モデルである HHMM に対応する識別モデル，階層隠れ CRF(HHCRF) を提案し，脳波データのラベル付け，セグメンテーションにてその有効性を確認した [1].

本研究では，HHCRF の新たなアプリケーションとして，バレーボールの試合映像のラベル付け，セグメンテーションを試みる．脳波データでは 2 層モデルを用いたが，本研究では 3 層モデルにより，スポーツ映像の繰り返し構造をモデル化する．

2 HHMM

HMM は音声認識においてよく用いられてきた時系列モデルである．そして，HHMM は HMM の状態空間に階層構造を持たせたモデルである．下層の状態では短期的な部分時系列を表現し，上層の状態ではより長期的な部分時系列を表現する [2].

HHMM は内部状態，出力状態，終了状態の 3 つの状態を持ち，縦遷移確率，横遷移確率，終了状態への遷移確率に従って状態遷移する．図 1(a) に階層数 3($D = 3$) の HHMM の DBN 表現を示す．図中の q_t^d ($d \in \{1,\dots,D\}$) は，時刻 t における深さ d の状態変数を表す．f_t^d は指標変数と呼び，q_t^d が時刻 t において終了状態に遷移する場合に 1，それ以外で 0 をとる．

3 HHCRF

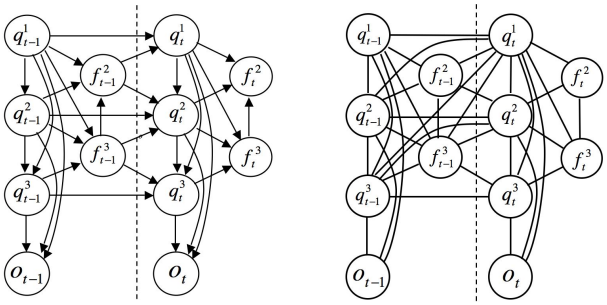
3.1 モデル

HHCRF は図 1(b) に示す無向グラフであり，T 個の観測値を持つ，ある時系列データ $O = \{o_1,\dots,o_T\}$ に対する，ある状態系列 $[Q^{1:U}, F^{2:U+1}]$ の条件付確率を以下のように表現する [2].

$$p(Q^{1:U}, F^{2:U+1}|O;\Lambda) = \frac{1}{Z(O;\Lambda)} \sum_{Q^{U+1:D}} \sum_{F^{U+2:D}} \cdot \exp\Big(\sum_{k=1}^{K} \lambda_k \Phi_k(Q^{1:D}, F^{2:D}, O) \Big)$$

ただし，$Q^{1:U} = \{q_1^{1:U},\dots,q_T^{1:U}\}_{U\in\{1,\dots,D\}}$ は上層の状態変数系列，$F^{2:U+1} = \{f_1^{2:U+1},\dots,f_T^{2:U+1}\}$ は上層の指標変数系列を表す．また，$\Phi_k(Q^{1:D}, F^{2:D}, O)$ は素性関数と呼び，λ_k は素性関数 Φ_k の重み

† 広島市立大学大学院情報科学研究科
〒731-3194 広島市安佐南区大塚東 3–4–1
Email: tamada@robotics.im.hiroshima-cu.ac.jp



(a) (b)

図 1 (a) HHMM の DBN 表現，(b) HHCRF の無向グラフ（共に時刻 $t-1$ から t に関する部分のみ）

である．ここで，モデルに含まれる全ての重み $\Lambda = \{\lambda_1,\dots,\lambda_K\}$ が HHCRF のモデルパラメータである．素性関数の選択は任意なので，モデルの性能を向上させる素性関数を自由に選択できる．分母の $Z(O;\Lambda)$ は正規化項である．

3.2 パラメータ推定

訓練データ集合 $\mathcal{D} = \{O^{(n)}, Q^{1:U(n)}, F^{2:U+1(n)}\}_{n=1}^{N}$ に対する最尤法に基づき，HHCRF のモデルパラメータを推定する．最適なモデルパラメータ $\hat{\Lambda}$ を推定するために必要となる，条件付き対数尤度 $\mathcal{L}(\Lambda) = \sum_{n=1}^{N} \log p(Q^{1:U(n)}, F^{2:U+1(n)}|O^{(n)};\Lambda)$ の偏微分は，以下のように与えられる．

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \lambda_k} = &\Bigg(\sum_{n=1}^{N} \sum_{Q^{U+1:D}} \sum_{F^{U+2:D}} \cdot \Phi_k(Q^{1:U(n)}, Q^{U+1:D}, F^{2:U+1(n)}, F^{U+2:D}, O^{(n)}) \\ &\cdot p(Q^{U+1:D}, F^{U+2:D}|Q^{1:U(n)}, F^{2:U+1(n)}, O^{(n)};\Lambda) \Bigg) \\ &- \Bigg(\sum_{n=1}^{N} \sum_{Q^{1:U}} \sum_{Q^{U+1:D}} \sum_{F^{2:U+1}} \sum_{F^{U+2:D}} \Phi_k(Q^{1:D}, F^{2:D}, O^{(n)}) \\ &\cdot p(Q^{1:D}, F^{2:D}|O^{(n)};\Lambda) \Bigg) \quad (1)\end{aligned}$$

方程式 (1) の右辺は，$p(Q^{U+1:D}, F^{U+2:D}|Q^{1:U(n)}, F^{2:U+1(n)}, O^{(n)};\Lambda)$ に対する素性の期待値と，$p(Q^{1:D}, F^{2:D}|O^{(n)};\Lambda)$ に対する素性の期待値の差を表す．

ここで，右辺第一項を計算するために必要な十分統計量は，$\{p(q_{t-1}^{U+1:D}, q_t^{U+1:D}, f_{t-1}^{U+2:D}, f_t^{U+2:D}|Q^{1:U(n)}, F^{2:U+1(n)}, O^{(n)};\Lambda)|1 \le t \le T\}$ と $\{p(q_t^{U+1:D}, f_t^{U+2:D}|Q^{1:U(n)}, F^{2:U+1(n)}, O^{(n)};\Lambda)|1 \le t \le T\}$ である．訓練で与えられる状態系列は上層のみの状態系列 $[Q^{1:U}, F^{2:U+1}]$ なので，これらの確率は，Backward-Forward-

Backward アルゴリズムにより求まる [3]．また，右辺第二項の期待値の十分統計量は，$\{p(q_{t-1}^{1:D}, q_t^{1:D}, f_{t-1}^{2:D}, f_t^{2:D}|O^{(n)}; \Lambda)|1 \leq t \leq T\}$ と $\{p(q_t^{1:D}, f_t^{2:D}|O^{(n)}; \Lambda)|1 \leq t \leq T\}$ であり，これらの確率は，Forward-Backward アルゴリズムにより求まる．

4 セグメンテーション

時系列データを意味のある区間へと分割することをセグメンテーションという．本研究では，まず，与えられた時系列データに対してラベル付けを行い，最適状態系列を推定する．次に，最適状態系列から各層における終了状態への遷移を検出し，その上の層のセグメント境界とする．最適状態系列の推定アルゴリズムには，Forward-Backward アルゴリズム (FB) と Viterbi アルゴリズムを用いる．

5 実験

同一のカメラが撮影した映像区間を「カット」，映像中のカメラの切換え点を「ショット」，動画の最小単位を「フレーム」と呼ぶ．スポーツのテレビ映像にはカット構成の規則性と階層構造が見られ，それらを利用することでスポーツ映像をプレイ単位の映像区間へと分割できる [4]．実験では，カット構成の規則性と階層構造を HHMM，HHCRF でモデル化してスポーツ映像のセグメンテーション実験を行い．HHCRF の優位性を示す．

5.1 時系列データと特徴抽出

本研究で扱うスポーツ映像は，CM が含まれないバレーボールのテレビ映像である．映像ファイルから時系列データを作成するために，まずフレームごとに各画素の特徴量を抽出した．特徴量は RGB の輝度値（3 次元）とオプティカルフロー（2 次元）である．それらの平均と分散を計算した 10 次元ベクトルを，更に主成分分析して得られる 4 次元ベクトルを時系列データとした．

5.2 モデル選択

サーブから始まり，どちらかのチームが得点するまでの映像区間をバレーボールにおけるプレイ単位の映像区間とする．その映像区間は典型的に以下の順のカット構成から成る．(1) サーブを構えた選手のカット，(2) ラリー中のコート全体を映したカット，(3) 点が決まった後のどちらか一方のチームのアップ，(4) リプレイを映したカット．映像全体にわたって，このカット構成が繰り返し現れる．我々はこの繰り返し構造を 3 層モデルによりモデル化した．モデルの第 1 層はプレイもしくはタイムアウトの 2 つの状態を表現し，第 2 層は上記のような 4 つのカットの状態を表現する．また，第 3 層はカット中のフレームの状態を表現するが，これは隠れ状態とした．

素性関数の選択については，状態遷移に関する素性関数は HHMM のモデル構造と等しくなるように選択した．また，HHMM は現時刻の観測値しか考慮しないが，観測に関する素性関数は，時刻をまたいで複数の観測値をとれるよう選択した．

5.3 ラベル付け実験

訓練データは 3 セット分の映像データとし，テストデータは訓練データとは異なる 1 セット分の映像データとした．訓練データを用いて学習した HHMM と HHCRF によって，テストデータに対するラベル付けを行い，最適状態系列を推定した．

実験を 10 回繰り返したときの，最適状態系列の正解率の平均値を表 1 に示す．ただし，各時刻において第 1 層および第 2 層の状態変数と第 2 層の指標変数の 3 つ全てが正しい場合を正解とした．表 1 より，HHCRF が HHMM よりも正確に最適状態系列を推定できていることが確認できる．

表 1　ラベル付け実験の結果

	HHMM		HHCRF	
	FB	Viterbi	FB	Viterbi
正解率 [%]	68.15	67.65	69.07	68.39

表 2　セグメンテーション実験の結果

	HHMM		HHCRF	
	FB	Viterbi	FB	Viterbi
適合率 [%]	26.80	23.85	28.41	23.76
再現率 [%]	38.32	53.27	43.77	54.34
F 値 [%]	31.54	32.95	34.45	33.07

5.4 セグメンテーション実験

ラベル付け実験で推定した最適状態系列から，第 2 層の指標変数 f_t^2 が 1 となるフレーム t を抽出し，第 1 層のセグメント境界の推定結果を得た．第 1 層のセグメント境界はプレイとプレイの間，またはプレイとタイムアウトの間のショットを表す．すなわち，第 1 層のセグメント境界は映像をプレイ単位へと分割する．

表 2 にセグメンテーション実験の結果を示す．セグメンテーションの性能は，適合率，再現率，F 値により評価する．表 2 より，セグメンテーションの性能についても，HHCRF が HHMM よりも勝っていることが確認できる．

6 まとめ

3 層構造の HHCRF を用いてスポーツ映像をモデル化し，ラベル付け実験，セグメンテーション実験において HHCRF の優位性を確認した．

HHCRF は素性関数を自由にとり得る．よって，有効な素性関数を更にモデルに追加すれば，性能を向上させることができると考えられる．

参考文献

[1] Sugiura, T., Gotou, N., Hayashi, A., ”Hierarchical Hidden Conditional Random Fields”, 第 21 回信号処理シンポジウム予稿集, pp. 628-633, 2007.

[2] K. P. Murphy, ”Linear Time Inference in Hierarchical HMMs”, Advances in Neural Information Processing Systems 14, vol. 2, pp. 833-848, 2002.

[3] Scheffer, T., Wrobel, S., ”Active hidden Markov models for information extraction” , Lecture Notes in Computer Science, vol. 2189, pp. 309-318, 2001.

[4] 椋木　雅之，寺尾　元宏，池田　克夫，”カット構成の規則性を利用したスポーツ映像のプレイ単位への分割”，電子情報通信学会論文誌, vol. J85-D-II, no. 6, pp. 1016-1024, 2002.