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

信学技報
TECHNICAL REPORT OF IEICE.

モーダル間の共起関係を考慮した 階層的トピック軌跡モデルによる映像認識検索

中野　拓帆† 　木村　昭悟†† 　亀岡　弘和†† 　宮部　滋樹† 　嵯峨山茂樹†
小野　順貴† 　柏野　邦夫†† 　西本　卓也†

† 東京大学 大学院情報理工学系研究科
†† 日本電信電話 (株) コミュニケーション科学基礎研究所
E-mail: †{t-nakano,miyabe,sagayama,onono,nishi}@hil.t.u-tokyo.ac.jp, ††akisato@ieee.org,
†††{kameoka,kunio}@eye.brl.ntt.co.jp

あらまし　本稿では，与えられた映像に適合するメタ情報を提示する映像認識（automatic video annotation）と，与えられたメタ情報に適合する映像を提示する映像検索（video retrieval）とを，統一的な枠組で取り扱う映像認識検索問題を取り上げ，そのための統計モデルである階層的トピック軌跡モデル HTTM を提案する．提案モデルは，各モダリティ及びクロスモーダルの共起関係を考慮したトピックモデルと，その時空間的ダイナミクスを表現する状態空間モデルとによって構成され，映像におけるインスタンス・シーン・コンセプトを階層的に表現する．このモデルに基づき，モデル推定・映像認識・映像検索それぞれを簡易に実現することが可能である．それと共に，音響信号や地理情報など他の要素を新規に導入する拡張も容易である．本稿では，人手によりラベル付けされた映像データセットに対してこのモデルを用いた認識実験を行い，精度向上の結果とともに報告する．
キーワード　映像認識・検索，トピックモデル，確率的正準相関分析，隠れマルコフモデル（HMM）

Hierarchical topic trajectory model for video annotation retrieval considering cross-modal co-occurrences

Takuho NAKANO†, Akisato KIMURA††, Hirokazu KAMEOKA††, Shigeki MIYABE†, Shigeki SAGAYAMA†, Nobutaka ONO†, Kunio KASHINO††, and Takuya NISHIMOTO†

† Graduate School of Information Science and Technology, The University of Tokyo
7–3–1 Hongo, Bunkyo-ward, Tokyo 113–8656 JAPAN
†† NTT Communication Science Laboratories, NTT Corporation
3–1 Morinosato Wakamiya, Atsugi, Kanagawa 243–0198 JAPAN
E-mail: †{t-nakano,miyabe,sagayama,onono,nishi}@hil.t.u-tokyo.ac.jp, ††akisato@ieee.org,
†††{kameoka,kunio}@eye.brl.ntt.co.jp

Abstract　This paper deals with a problem of “video annotation retrieval” that achieves automatic video annotaion (providing relevant text labels for a given video clip) and video retrieval (providing relevant video clips for a given text label(s)) within a unified framework. We propose a new statistical model, named Hierarchical Topic Trajectory Model (HTTM), for aquiring a dynamically changing topic model that represents the relationship between video frames and associated text labels. Model parameter estimation, annotaion and retrieval can be easily executed. It is also easy to add new modals such as audio signal and geometrical information. Preliminary experiments on video annotation task with humanly annotated video dataset indicate that our proposed method can improve the annotation accuracy.
Key words　video annotation retrieval, topic model, canonical correlation analysis, hidden Markov model(HMM)

1. はじめに

テレビやラジオ放送のデジタル化，WEB を介した音楽・映像の配信・共有サービスの普及により，音・映像のデータは爆発的な勢いで増加し続けており，こういった大量かつ大容量の時系列データに対する認識・検索技術に対する社会的な需要や期待が高まっている．

静止画像に含まれる物体の意味カテゴリを認識する問題は，

一般物体認識と呼ばれている．これにならい，写真や映像の中の一般的な意味カテゴリを認識する問題を，以降,（一般）画像認識と呼ぶことにする．一般画像認識によって，未知画像に対して自動的に適切なテキストラベルを付与する「認識」課題と，テキストや画像をクエリとし，画像を提示する「検索」課題が可能となることが期待され，盛んに研究されている [1].

画像認識は対象を限定しないばかりか，場合によっては,「遠足」,「夕焼け」といった，直接対応する物体がない場合も考慮することがあるため，非常に難しい課題として知られている [2], [3]．最近は，人手付けされたデータベース [4], [5] や，Web データ [6] を用いて機械学習を行い，未知対象に認識を行う研究が数多く行われている [7]．本研究もこの流れに即した統計的機械学習モデルに基づくアプローチを採る．

機械学習を画像認識に適用する際の標準的な方法として，各ラベルに対応する one-versus-all の (2 値) 識別器を．support vector machine (SVM) [8] 等を用いて構成し，判定する方法が考えられる，しかし，これらの手法にはいくつかの問題点がある．まず，複数ラベルに対する効率的な識別機の設計が困難であるという問題である．道路，車，バイク，横断歩道といったラベルには互いに関連があると考えられるが，2 値識別器に基づく手法ではこれらの相関を明示的に取り扱うために数多くの識別器が必要となる．また，これらの手法には学習における計算量や記憶容量が多大になるものも多い．

これに対し，画像やラベルの上位に潜在空間，潜在変数を仮定してモデル化する生成的アプローチも知られている．このようなモデルはトピックモデルと呼ばれることもある [9], [10]．代表的なトピックモデルとして，probabilistic latent semantic analysis (pLSA) [11], [12] や latent Dirichlet allocation (LDA) [13], [14] などが知られる．トピックモデルの簡易導出法のひとつとして，確率的正準相関分析（probabilistic canonical correlation analysis; PCCA）[15], [16] をベースにした手法があり [17], [18]，良好な認識結果が得られている．PCCA は多変量解析に基づいてモデルが構成されるため動作が高速なことに加え，カーネル法を用いた非線形拡張も容易であり [19], [20]，今後の応用可能性が高いと思われる．本稿でも潜在変数を用いる手法を採用する．

一方で，パターン認識的な側面から映像の認識・検索について取り組まれている研究を見直すと，画像情報に特化して解析を行うものが多い．その標準的な手順は，まず，映像をショット（shot）と呼ばれる小単位に分割した上で，ショット内の画像を取り出し，取り出された画像について，画像認識・検索と同じ枠組みで処理を行うものである．しかし，これらのアプローチは，映像の特徴をそのモデルに活かしきれていないと考えられる．以下その大きな理由を 2 つ述べる．

第一の理由は，映像は画像だけでなく音響情報や，ニュース原稿，さらに，付与された撮影時刻，場所，放送・録画時刻という多様な情報を含む，マルチモーダルデータであるということである．この多様性への対応として，自動音声認識 (automatic speech recognition; ASR) 結果が特徴量として準公式的に提供されている他 [21]，音響特徴の利用により，特に singing というラベルについて大幅な認識結果の向上 [22]，画像認識による位置情報利用の検討 [23] が報告されている．

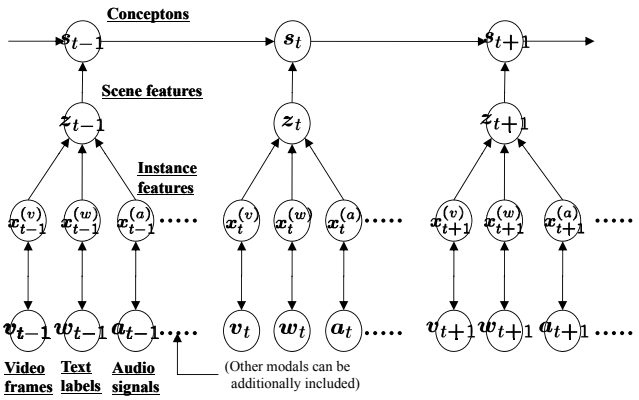

図 1　提案モデル HTTM の概念図．変数同士は矢印で互いに関係している．$\boldsymbol{x}_t^{(v)}, \boldsymbol{x}_t^{(w)}, \boldsymbol{x}_t^{(a)}$ は各インスタンス v_t, w_t, a_t から抽出された特徴量であり，各インスタンスの性質を示す．潜在変数 $\boldsymbol{z}_t$ は，複数要素の特徴量にクロスモーダルに関わり，シーンのような概念を表していると考えられる．状態変数 s_t は潜在変数分布の時間的遷移を特徴づけ，コンセプトのような概念を表していると考えられる．

第二の理由は，映像は時系列データであるというという点である．映像は物体の細かい動きから，CM 切り替えなどの場面転換，起承転結といった大きな流れを合わせもっており，それらも適切にモデル化されることが望ましい．時系列性を考慮した先行研究として，複数フレームの利用による認識率向上 [24] や，映像特徴とテキスト，音響特徴との関係を階層化させ，中間階層のグルーピングを階層型 HMM や pLSA でモデル化し，教師なし学習によってニュース映像を自動分類する研究 [25] などが知られている．

これらの状況を鑑み，本稿では映像のマルチモーダル性と時系列性を考慮した新しい統計モデル，階層的トピック軌跡モデル (hirarchical topic trajectory model; HTTM) を提案する．

以下，本稿ではまず第 2. 節で提案モデルの概要を示す．次に第 3. 節でモデル学習手法を，第 4. 節でモデルを用いた認識手法を示す．第 5. 節では映像の認識実験により提案手法の有効性を示し，第 6. 節で結論と今後の展望について述べる．

2. 提案モデル概要

提案モデル HTTM の概念図を図 1 に示す．最下層の観測として，実信号に対応する，映像フレーム，テキストラベル，音響信号 (v_t, w_t, a_t) などがあり．次の層として，特徴量 $\boldsymbol{x}_t^{(v)}, \boldsymbol{x}_t^{(w)}, \boldsymbol{x}_t^{(a)}$ が各信号からの特徴量抽出手法で得られる．その次の層に，クロスモーダルの共起関係反映する潜在変数 $\boldsymbol{z}_t$ を，映像ショット毎に仮定する．最上層として，潜在変数を出力とする隠れ状態の系列 s_t を仮定する．

このようにモデル化することで，一般的なトピックモデルにおける「トピック」に相当する潜在変数 $\boldsymbol{z}_t$ が時間的に遷移する状況を記述することが可能になる．概念的には潜在変数 $\boldsymbol{z}_t$ によってショットのシーンを特徴づけ，状態変数 s_t の時系列に

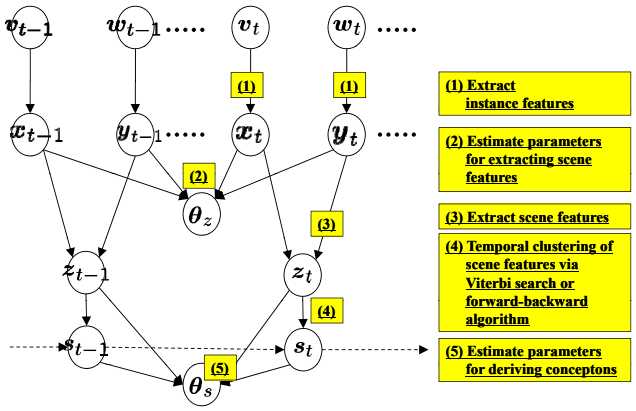

図 2　HTTM の統計モデル学習フロー.

よって，当該映像空間のコンセプトを特徴付けると考えることが可能である.

以降，HTTM の詳細について述べる．各時刻 t における状態変数 s_t は K 個の離散的な状態変数集合 $\boldsymbol{S}(=\{1,\ldots,K\})$ のいずれかの値をとる．時刻 t での状態が $s_t = i$ であったとき，時刻 $t+1$ での状態が $s_{t+1} = j$ となる確率を状態遷移確率 p_{ij} とする．潜在変数 $\boldsymbol{z}$ の出力確率密度関数は，以下の混合ガウス分布 (Gaussian mixture model; GMM) によりモデル化さる.

$$p(\boldsymbol{z}_t|s_t = k) = \sum_{j=1}^{L_k} \pi_{k,j} p(\boldsymbol{z}_t|j, s_t = k), \qquad (1)$$
$$p(\boldsymbol{z}_t|j, s_t = k) = \mathcal{N}(\boldsymbol{z}_t; \bar{\boldsymbol{z}}_{k,j}, \Sigma_{k,j})$$

ここで，L_k は状態 k に対応する GMM の混合数，$\bar{\boldsymbol{z}}_{i,j}, \Sigma_{k,j}, \pi_{k,j}$ はそれぞれ j 番目の正規分布の平均ベクトル，分散共分散行列，重みであり，$\mathcal{N}(\cdot; \boldsymbol{\mu}, \Sigma)$ は平均 $\boldsymbol{\mu}$，分散共分散行列 Σ の正規分布を示す．各モダリティの実信号 v_t, w_t, a_t に対応する特徴ベクトル $\boldsymbol{x}_t^{(v)}, \boldsymbol{x}_t^{(w)}, \boldsymbol{x}_t^{(a)}$ は，潜在変数 $\boldsymbol{z}_t$ に依存した分布から生成される確率変数となることを仮定する．例えば，映像特徴ベクトル $\boldsymbol{x}_t^{(v)}$ は潜在変数 $\boldsymbol{z}_t$ のアフィン変換を平均とする下記の正規分布に従う確率変数であると仮定する:

$$p(\boldsymbol{x}_t^{(v)}|\boldsymbol{z}_t) = \mathcal{N}(\boldsymbol{x}_t^{(v)}; W_v \boldsymbol{z}_t + \bar{x}^{(v)}, \Psi_v).$$

他の特徴ベクトルについても同様にモデル化する.

3.　HTTM のモデル学習

3.1　パラメタ学習の概要

パラメタ学習のフローを図 2 に示す．以下，説明や記述を簡単にするため，画像特徴とラベル特徴の 2 要素についてのみ考慮し，$(\boldsymbol{x}_t^{(v)}, \boldsymbol{x}_t^{(w)})$ の代わりに $(\boldsymbol{x}, \boldsymbol{y})$ を用いる（$\boldsymbol{x}_t^{(a)}$ については省略する）．まず各モダリティの特徴ベクトル $\boldsymbol{x}, \boldsymbol{y}$ の間の関係を分析することにより，潜在変数 $\boldsymbol{z}$ が存在する部分空間を導出すると共に，各時刻の潜在変数 $\boldsymbol{z}_t$ を求める．次に潜在変数 $\boldsymbol{z}_t$ を観測とする HMM を学習する．以降，潜在変数空間を表現するモデルパラメタの集合を $\boldsymbol{\theta}_z$，HMM のモデルパラメタ集合を $\boldsymbol{\theta}_s$ とおく.

3.2　潜在変数空間の獲得と潜在変数の推定

本節では，潜在空間の獲得と潜在変数の推定について述べる(注1)．本稿では，計算の簡易さと応用可能性を考え，潜在空間の獲得に確率的正準相関分析（PCCA）[15] を用いた.

p 次元の画像特徴を $\boldsymbol{x} = (x_1, x_2, \ldots, x_p)^\top$，$q$ 次元のラベル特徴を $\boldsymbol{y} = (y_1, y_2, \ldots, y_q)^\top$，潜在空間の次元を d とおく．ここで $d \leqq \min(p, q)$ である．画像の画像特徴 $\boldsymbol{x}$ のみが与えられた場合，画像特徴 $\boldsymbol{x}$ とラベル特徴 $\boldsymbol{y}$ の両方が与えられた場合それぞれについて，潜在変数 $\boldsymbol{z}$ が以下の形で与えられることが知られている [15], [28].

$$\boldsymbol{z}(\boldsymbol{x}) = M_x^\top A^\top (\boldsymbol{x} - \bar{\boldsymbol{x}}), \qquad (2)$$

$$\boldsymbol{z}(\boldsymbol{x}, \boldsymbol{y}) = \begin{pmatrix} M_x \\ M_y \end{pmatrix}^\top \begin{pmatrix} E & -E\Lambda \\ -E\Lambda & E \end{pmatrix} \begin{pmatrix} A^\top(\boldsymbol{x} - \bar{\boldsymbol{x}}) \\ B^\top(\boldsymbol{y} - \bar{\boldsymbol{y}}) \end{pmatrix}, \qquad (3)$$

$$E = (I - \Lambda^2)^{-1}$$

上記の式に含まれる，潜在変数空間を規定するモデルパラメタ集合 $\boldsymbol{\theta}_z = \{M_x, M_y, \Lambda, A, B, \bar{\boldsymbol{x}}, \bar{\boldsymbol{y}}\}$ を推定する.

まず，$\bar{\boldsymbol{x}}, \bar{\boldsymbol{y}}$ は学習用データの平均をとることで推定できる．残りのパラメタについては，正準相関分析と同様の手順で得られる．まず，以下に示す一般化固有値問題を解き，固有値の大きい順に d 個の固有値と固有ベクトル $\lambda_i, \boldsymbol{a}_i \in \mathbf{R}^p, \boldsymbol{b}_i \in \mathbf{R}^q (i = 1, 2, \ldots d)$ を得る.

$$\begin{pmatrix} \mathbf{0} & C_{xy} \\ C_{yx} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \boldsymbol{a} \\ \boldsymbol{b} \end{pmatrix} = \lambda \begin{pmatrix} C_{xx} & \mathbf{0} \\ \mathbf{0} & C_{yy} \end{pmatrix} \begin{pmatrix} \boldsymbol{a} \\ \boldsymbol{b} \end{pmatrix},$$

ここで，$\mathbf{0}$ は零行列，$C_{xx}, C_{yy}, C_{xy}, C_{yx}$ は（学習用データから得られる）分散共分散行列とする．次に，λ_i を大きい順に d 個対角に並べた対角行列として Λ を獲得する．また，固有値に対応する固有ベクトル（列ベクトル）を行方向に並べることで $A = (\boldsymbol{a}_1, \ldots, \boldsymbol{a}_d), B = (\boldsymbol{b}_1, \ldots, \boldsymbol{b}_d)$ を得る．また $M_x, M_y \in \mathbf{R}^{d \times d}$ は，$M_x M_y^\top = \Lambda$ かつスペクトルノルムがそれぞれ 1 未満という条件を満たす任意の行列とする.

3.3　HMM モデルパラメタの推定

潜在変数と状態変数の間のモデルとして，潜在変数を観測，状態変数を隠れ状態，混合ガウス分布 (GMM) を出力確率とする HMM を利用した．簡単のため，各状態の GMM の混合数を状態変数によらず一定，つまり $L_k = L$ とする．Baum-Welch アルゴリズム [29] や Viterbi 学習 [30] など，一般的な HMM の学習法を用いて，状態遷移確率及び，出力確率のためのパラメタである平均ベクトルと分散共分散行列，混合重みの学習を行ってモデルパラメタ集合 $\boldsymbol{\theta}_s = \{\{p_{ij}\}, \{\bar{\boldsymbol{z}}_{k,j}, \Sigma_{k,j}, \pi_{k,j}\}\}$ を得る.

4.　HTTM を用いた映像の認識

4.1　欠損特徴量の推定

本節では，提案モデル HTTM を用いた映像認識検索の具体的なアルゴリズムについて述べる.

認識課題の場合，映像データが与えられ，ラベルのデータは与えられない．このとき，映像特徴からラベル（言語）情報を

(注1)：正準相関分析の 3 変量以上の場合への拡張は [26], [27] を参照.

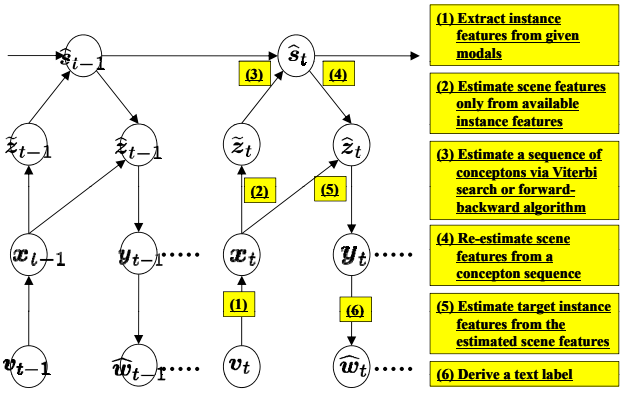

図 3　HTTM の認識フロー.

推定することで認識が可能になる．図 3 は認識課題を例にした，HTTM におけるラベル推定の概要である．一方，検索課題の場合，ラベルのデータが与えられ，映像のデータは与えられない．この時，ラベル情報から映像情報を推定し，データベースから適合する映像を提示することになる．検索・認識いずれの場合でも，一部の情報が欠損しており，欠損部分の特徴を推定するという点で同じであり，認識と検索が対称的な同一の枠組で行うことが可能である．以降，認識課題の手順を説明する．検索課題も同様に行える．

(1)　まず，画像情報から特徴抽出を行う．

(2)　抽出した画像特徴量 $\boldsymbol{x}_t$ から，潜在変数 $\tilde{\boldsymbol{z}}_t = \boldsymbol{z}(\boldsymbol{x}_t)$ を推定する．このとき，モデルパラメタ集合はモデル学習で得られた $\boldsymbol{\theta}_z$ を用いる．

(3)　モデル学習で得られたモデルパラメタ集合 $\boldsymbol{\theta}_s$ を用い，潜在変数の系列 $\tilde{\boldsymbol{Z}} = \{\tilde{\boldsymbol{z}}_1, \tilde{\boldsymbol{z}}_2, \ldots, \tilde{\boldsymbol{z}}_t, \ldots\}$ から，状態変数の系列 $\hat{\boldsymbol{S}} = \{\hat{s}_1, \hat{s}_2, \ldots, \hat{s}_t, \ldots\}$ を推定する．状態変数の推定は，もっとも簡単な場合，Viterbi decoding によって得られる:

$$\hat{\boldsymbol{S}} \approx \underset{S}{\operatorname{argmax}}\, P(\tilde{\boldsymbol{z}}_1|s_1)P(s_1)\prod_{t=2}^{T} P(\tilde{\boldsymbol{z}}_t|s_t)P(s_t|s_{t-1}).$$

ただし，T は与えられた映像のショット数である．Viterbi decoding に代えて，forward-backword アルゴリズムを用いて，状態変数系列 $\hat{S}$ 確率的に推定することもできる．

(4)　推定された状態変数 $\hat{s}_t$ を用いて，潜在変数を再推定する．再推定された潜在変数を $\hat{\boldsymbol{z}}_t$ と書く．推定の際には，状態 $\hat{s}_t$ に対応する GMM パラメタ集合（平均ベクトル，分散共分散行列，重み $\bar{\boldsymbol{z}}_i, \Sigma_i, \pi_i (i = 1, 2, \ldots, L)$）と，画像特徴量 $\boldsymbol{x}_t$ だけから推定された潜在変数 $\tilde{\boldsymbol{z}}_t$ が利用できる．$\hat{\boldsymbol{z}}_t$ は以下の式で計算される：

$$\hat{\boldsymbol{z}}_t = \sum_{i=1}^{L} \tilde{\pi}_i \bar{\boldsymbol{z}}_i.$$

ただし，$\tilde{\pi}_i$ は以下で与えられる：

$$\tilde{\pi}_i = \frac{\pi_i \mathcal{N}(\tilde{\boldsymbol{z}}_t; \bar{\boldsymbol{z}}_i, \Sigma_i)}{\sum_{j=1}^{L} \pi_j \mathcal{N}(\tilde{\boldsymbol{z}}_t; \bar{\boldsymbol{z}}_j, \Sigma_j)}.$$

(5)　再推定された潜在変数 $\hat{\boldsymbol{z}}_t$ と $\boldsymbol{x}_t$ を用いて，ラベル特徴量 $\hat{\boldsymbol{y}}$ を推定する．PCCA の原理から，

$$\hat{\boldsymbol{y}} = \boldsymbol{y}(\boldsymbol{z}) = W_y \boldsymbol{z} + \bar{\boldsymbol{y}} \tag{4}$$

となる．ただし，W_y は以下の式で与えられる：

$$W_y = C_{yy} B M_y$$

M_y については，対角行列 $M_y = \Lambda^{1-\beta} (\beta = 0.5)$ で与える．

(6)　推定されたラベル特徴量 $\hat{\boldsymbol{x}}_t$ を用い，ラベル推定を行う．これについては次節にて述べる．

4.2　特徴量からの情報の復元・検索

認識課題では，前節の方法で得られたラベル特徴からラベル情報を推定・復元する．推定・復元方法は，ラベル特徴の種類に依存するので，ラベル特徴の部分で述べる．最終的な認識結果としては，一般的な方法と同じく，

- 各画像について物体認識： ショットを固定して，ラベルについて，確率が高いと考えられる順番に付与判定を行う
- ラベルについて，付与画像を判定： ラベルを固定して，付与確率が高いと考えられる順にショットを出力する

ことなどを行えばよい．

検索課題では，データベース内の映像から取り出した特徴と，前節の方法で得られた映像特徴とを照合し，類似する特徴を持つ映像を提示することで行うことができる．

5.　映像の認識実験

5.1　実 験 条 件

提案モデルを用いて，公開されている画像特徴，ラベルのデータベースを用いて，認識実験を行った．

画像特徴は vireo374 [31] で公開されている局所特徴量から，500 次元の正規化された SIFT の BoF 特徴を用いた．

ラベルは，LSCOM [32] 及び LSCOM-Lite [33] を用いた．LSCOM は，TRECVID [5] の 2005 年のニュース映像データについて，人手で 449 種類の高次特徴ラベルについて，各ショットに対して付与の有無を記したものである．また，LSCOM-Lite は LSCOM と一部重複する 39 種類のラベルについて，同様のデータを公開している．今回は LSCOM-Lite に含まれる 39 種類のラベルに LSCOM のうち 8 種類のラベルを加えた，計 47 ラベルを学習用データとして用いた(注2)．

利用可能なデータのうち，47 ラベルの全てについて，有無の付与（判定）がされている 127 本分の映像，合計 56191 ショット分を用いた．さらにこれを約 4:1 に分割し，102 本，45689 ショットを学習用データ，25 本，10502 ショットをテスト用デー

(注2)：Airplane, Airplane_Flying, Animal, Boat_Ship, Building, Bus, Car, Charts, Cityscape, Classroom, Computer_TV-screen, Corporate-Leader, Court, Crowd, Demonstration_Or_Protest, Desert, Entertainment, Explosion_Fire, Face, Flag-US, Government-Leader, Hand, Maps, Meeting, Military, Mountain, Natural-Disaster, Nighttime, Office, Outdoor, People-Marching, Person, Police_Security, Prisoner, Road, Singing, Sky, Snow, Sports, Studio, Telephones, Truck, Urban, Vegetation, Walking_Running, Waterscape_Waterfront, Weather である．

タとした．

5.2 ラベル（言語）特徴の検討

トピックモデルとの組み合わせにおいて，ラベルの特徴をどのようにとるのが望ましいかは自明とは言えない．ラベル「有，無」に対応付けて「負，正」といった値を与える手法が有効な可能性もあるし，それぞれ 1, 0 の 2 値で与える手法も考えられる．．また，ある画像に対するラベル付与の有無を，ある文書に対する単語の有無のアナロジーとして考えることも可能であり，自然言語処理の手法が有効である可能性がある．

そのため，例えば以下のような手法の違いが認識にどのような影響を及ぼすかについて検証を行った．

- ラベル特徴を $\{0, 1\}$ のように，0，正の値の形で与える手法と $\{-1, 1\}$ のように，負の値，正の値の形で与える手法
- ラベル特徴に tf-idf 重み付けを行う手法と，そのままの値を用いる手法
- ラベル集合を小さな言語に見立て，潜在意味解析 (latent semantic analysis; LSA) を適用し，特徴抽出を行う手法と，そのままの値を用いる手法

そこで，以上 3 項目について，全ての組み合わせである 8 通りの組み合わせを試行した．ここで，ラベルに tf-idf 重み付けを行う手法と負の値，正の値の形で与える手法を組み合わせた場合，l 番目のラベルについて，付与がされている場合 $+\mathrm{idf}(l)$ されていない場合，$-\mathrm{idf}(l)$ の値を与えた．

5.3 評価手法

認識結果の評価は TRECVID における評価基準と同様に，ラベルに対する average precision (AP) と，そのラベル間平均である mean average precision (meanAP) により行う．ランク付け出力からの AP の計算式は

$$\mathrm{AP} = \frac{1}{R}\sum_{k=1}^{N} r_k \mathrm{precision}(k)$$

で与えられる．ここで，r_k は k 番目の出力結果が正解ならば $r_k = 1$，そうでなければ $r_k = 0$ となる数であり，$\mathrm{precision}(k)$ は 1 番目から k 番目までの正解率である．

仮にランダムに出力を行った場合，出力の各ランクにおける正解率は chance level 程度になる．よって，ランダム出力の場合，その加算平均をとる average precision も同じくらいの値になると考えられる．このように，chance level は average precision に対する良い指標となっている．そのため，ラベル別の実験結果には chance level を付記した．

5.4 実験結果

提案手法の出力結果は 4.2 節に記述した，ラベルについて，付与画像を判定する手法を用いた．つまり，l 番目のラベルの認識結果は，各時刻 t について推定されたラベル特徴量 $\hat{\boldsymbol{x}}_t$ の l 次元目の値を比較し，その値が大きい順に時刻 t をランク付け出力するものとした．

まず，潜在変数の次元を 47 に固定し，HMM の状態数及び混合数を変化させながら，ラベル特徴の取り方を変えた実験を行った．図 4 にその結果を示す．この図から，ラベル特徴を tf-idf 重み付けして，0 と正の値となるように与えた場合，もっとも meanAP の値が良くなることが確認できた．また，いずれの手法も画像データのみから推定した潜在変数 $\hat{z}_t$ を用いるよりも，提案手法 HTTM で再推定した潜在変数 $\hat{\boldsymbol{z}}_t$ を用いた方が多くのラベルで認識結果が向上していることが確認できた．

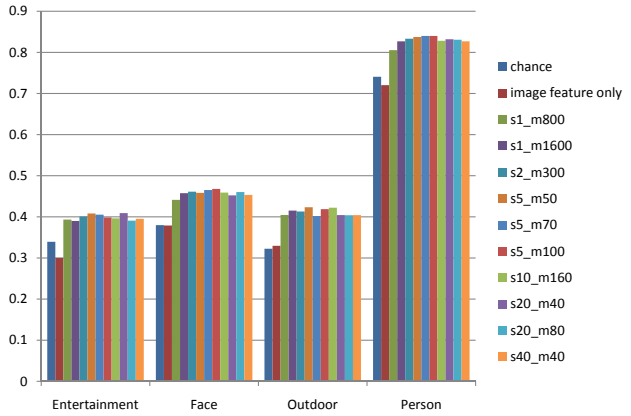

図 5　Entertainment, Face, Outdoor, Person の認識結果．比較的高い認識精度が得られている．

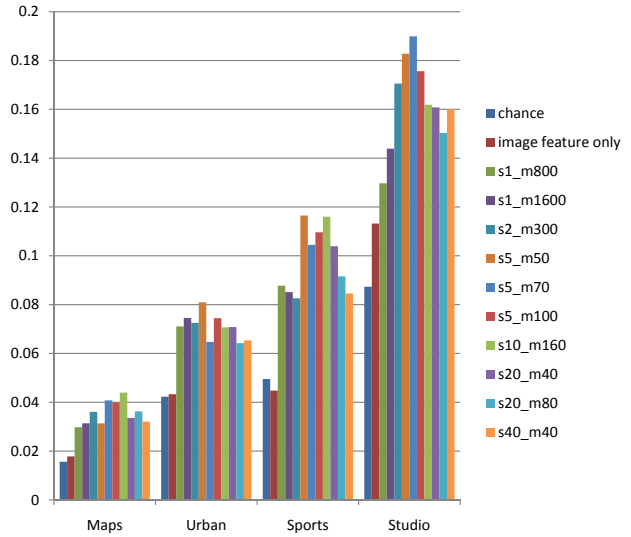

図 6　Maps, Urban, Sports, Studio の認識結果．認識結果の average precision と chance level に相関がみられる．

図 4 の結果に基づき，以降の実験ではラベル特徴を tf-idf 重み付けして，0 と正の値となるように与えた場合について，ラベル毎の認識結果についての分析を行った．

図 5 に他のラベルと比較して chance level が高いという特徴を持っている Entertainment, Face, Outdoor, Person の認識結果を示した．画像個別に，その特徴だけから認識するのに比べ，提案手法での結果向上が確認できた．また，全体的な傾向として，認識結果の average precision と chance level には相関がみられることが確認できた．図 6 の Maps, Urban, Sports, Studio の認識結果がその例である．chance level は学習用データのラベル頻度と類似した値をとるため，学習データが十分にあれば認識性能が向上することが示唆される．また，Studio ラベルについては状態数 5 のものが特に結果が良い，これは，

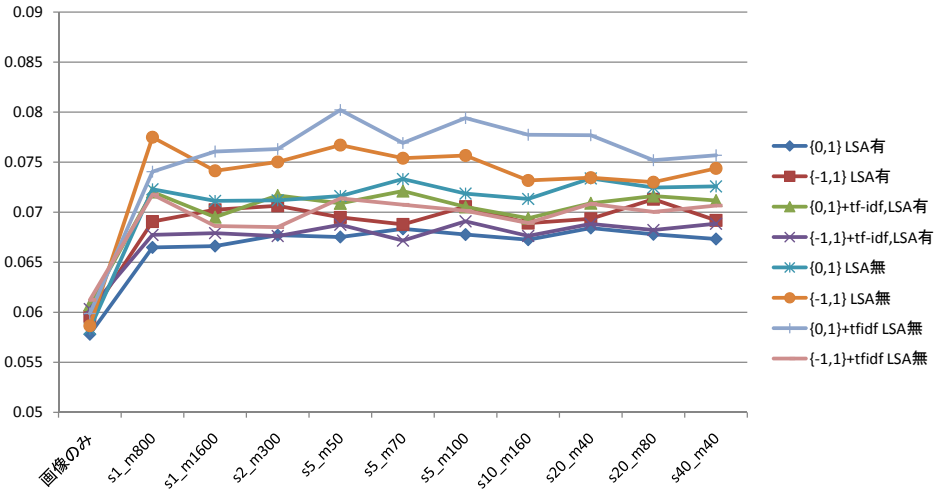

図 4　テスト用データに対して認識を行った結果．縦軸は meanAP である．「画像のみ」は，提案手法を用いず，画像のみから認識を行った結果である．それ以外は，$sn_s\_mn_m$ という形式で，n_x に HMM の状態数，n_m に各状態における GMM 混合数を示した．

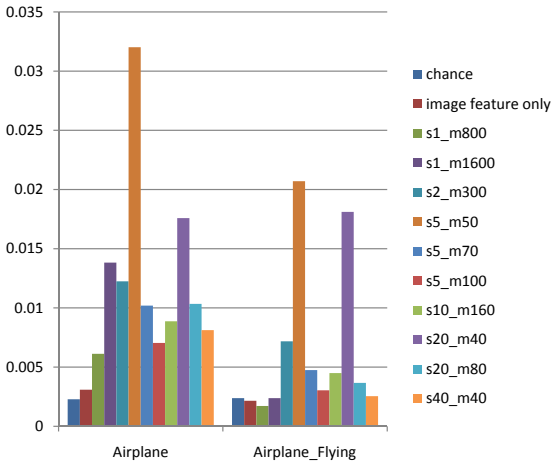

図 7　Airplane, Airplane_Flying の認識結果．互いの認識結果に相関がみられる．

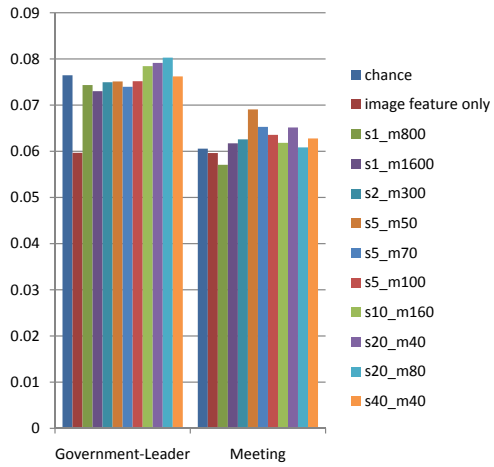

図 8　Government-Leader, Meeting の認識結果．結果は chance level と大差なく，認識できていないと考えられる．

ニュース映像おいて，studio が写っているシーンがある状態にまとまり，トピック遷移がうまくモデル化できていると考えられる．

図 7 は Airplane 及び Airplane_Flying の認識結果である．互いの認識結果に相関がみられ，トピック抽出を試みる提案手法の有効性を示している．

図 8 は Government-Leader 及び Meeting の認識結果である．結果は chance level と大差がない．chance level 程度の average precision 値であるため，前述の理由によって，これらラベルに関しては，正常に認識できていないと考えられる．

図 9 は Bus 及び Military の認識結果である．画像のみを用いて認識した結果と比較して，大きく結果が悪化している．どちらも chance level は非常に低く，GMM の学習が正常にされなかった可能性が考えられる．また全体として，状態数 1 混合数 800 の結果と 1600 の結果に大差がないものが多いが，一部

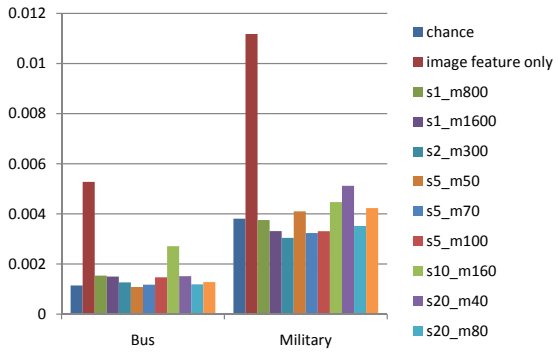

図 9　Bus, Military の認識結果．画像特徴だけからのラベル推定より結果が悪化している．HMM 部分がうまく機能していないと考えられる．

ラベルについては大きく結果が変化している．混合数を増やしすぎることでモデル過適合が起きている可能性が考えられる．

6. まとめと今後の方針

本稿では，映像のマルチモーダル性と時系列性を考慮した統計モデル，階層的トピック軌跡モデル (hirarchical topic trajectory model; HTTM) を提案した．

今回，TRECVID2005 の公開データベースに対して提案手法を適用し，提案法の有効性を示した．さらに，ラベル特徴の与え方について検討し，ラベル特徴を tf-idf 重み付けして，0 と正の値となるように与えた場合，もっとも認識結果が良くなることが確認された．

学習データ量は十分とは言えないため，学習データを増やして更なる検証を行いたい．それによって，一部ラベルの認識結果が改善する可能性もあると考えられる．

今回は潜在変数の次元数，HMM の状態数，出力確率の混合正規分布の混合数など，各種パラメタのいくつかの値を手動で与えたが，交差検定やパラメタの自動最適化などについては今後検討していく必要があると思われる．

謝辞 研究資源等の面からご支援頂いた，NTT コミュニケーション科学基礎研究所，上田修功 所長，前田英作 主席研究員，納谷太主幹研究員，大和淳司 部長に感謝する．また，Yu-Gang Jiang をはじめ，vireo374 [31] データセットの公開に関わった方々にもこの場を借りて謝意を表する．

文　献

[1] 柳井啓司，“[チュートリアル] 一般物体認識，” 電子情報通信学会研究会報告: パターン認識・メディア理解研究会，第 109 巻，pp.89–96，2009.

[2] 内田誠一，佐藤真一，鷲見和彦，福井和広，“パターン認識・メディア理解の問題分析，” 電子情報通信学会誌，vol.92，no.8，pp.656–664，2009.

[3] 鷲見和彦，内田誠一，佐藤真一，佐藤洋一，日浦慎作，福井和広，馬場口登，“パターン認識・メディア理解の 10 大チャレンジテーマ，” 電子情報通信学会誌，vol.92，no.8，pp.665–675，2009.

[4] M. Everingham, L. Van Gool, C.K.I. Williams, J. Winn, and A. Zisserman, “The pascal visual object classes (voc) challenge,” International Journal of Computer Vision, vol.88, no.2, pp.303–338, June 2010.

[5] A.F. Smeaton, P. Over, and W. Kraaij, “Evaluation campaigns and trecvid,” MIR ’06: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval, pp.321–330, ACM Press, New York, NY, USA, 2006.

[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “ImageNet: A Large-Scale Hierarchical Image Database,” CVPR09, pp.248–255, 2009.

[7] L. Fei-Fei, R. Fergus, and A. Torralba, “Recognizing and learning object categories: Year 2009,” ICCV2009 tutorial, pp.●●–●●, 2009.

[8] V.N. Vapnik, Statistical Learning Theory, Wiley-Interscience, Sept. 1998.

[9] T. Iwata, T. Yamada, and N. Ueda, “Modeling social annotation data with content relevance using a topic model,” Advances in Neural Information Processing Systems 22, eds. by Y. Bengio, D. Schuurmans, J. Lafferty, C.K.I. Williams, and A. Culotta, pp.835–843 ●●, , 2009.

[10] T. Iwata, T. Yamada, and N. Ueda, “Probabilistic latent semantic visualization: topic model for visualizing documents,” KDD ’08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, pp.363–371, ACM, New York, NY, USA, 2008.

[11] T. Hofmann, “Unsupervised learning by probabilistic latent semantic analysis,” Mach. Learn., vol.42, no.1-2, pp.177–196, 2001.

[12] K. Barnard and D. Forsyth, “Learning the semantics of words and pictures,” ●●, vol.2, pp.408–415vol.2, 2001.

[13] D.M. Blei, A.Y. Ng, and M.I. Jordan, “Latent dirichlet allocation,” J. Mach. Learn. Res., vol.3, pp.993–1022, 2003.

[14] L. Fei-Fei and P. Perona, “A bayesian hierarchical model for learning natural scene categories,” ●●, vol.2, pp.524–531vol.2, jun 2005.

[15] F.R. Bach and M.I. Jordan, “A probabilistic interpretation of canonical correlation analysis,” Technical report, Technical Report 688, Department of Statistics, University of California, Berkeley, 2005.

[16] C. Wang, “Variational bayesian approach to canonical correlation analysis,” Neural Networks, IEEE Transactions on, vol.18, no.3, pp.905–910, 2007.

[17] 中山英樹，原田達也，園吉康夫，“大規模 web 画像のための画像アノテーション・リトリーバル手法，” 画像の認識・理解シンポジウム（MIRU）2009，pp.103–110，2009.

[18] 木村昭悟，中野拓帆，杉山　将，亀岡弘和，前田英作，坂野　鋭，“SSCDE： 画像認識検索のための半教師付正準密度推定法，” 画像の認識・理解シンポジウム（MIRU）2010，pp.●●–●●，2010.

[19] M. Welling, “Kernel canonical correlation analysis,” ●●, pp.●●–●●, 2005.

[20] S. Akaho, “A kernel method for canonical correlation analysis,” In Proceedings of the International Meeting of the Psychometric Society (IMPS2001), pp.●●–●●, Springer-Verlag, 2001.

[21] “TRECVID data availability,” `http://trecvid.nist.gov/trecvid.data.html`.

[22] 井上中順，斉藤辰彦，篠田浩一，古井貞熙，“SIFT 混合ガウス分布と音響特徴を用いた映像からの高次特徴検出，” 電子情報通信学会研究会報告: パターン認識・メディア理解研究会，第 109 巻，pp.97–102，2009.

[23] 八重樫恵太，丸山拓馬，柳井啓司，“ジオタグ画像認識における位置情報の利用法の検討と分析，” 画像の認識・理解シンポジウム（MIRU）2010，pp.●●–●●，2010.

[24] C.G.M. Snoek, O.D. Rooij, B. Huurnink, J.C.V. Gemert, J.R.R. Uijlings, J. He, X. Li, I. Everts, V. Nedovic, M.V. Liempt, R.V. Balen, F. Yan, M.A. Tahir, K. Mikolajczyk, J. Kittler, M.D. Rijke, J.M. Geusebroek, Th. Gevers, M. Worring, A.W.M. Smeulders, and D.C. Koelma, “The MediaMill TRECVID 2008 Semantic Video Search Engine Draft notebook paper,” NIST TRECVID Workshop, pp.●●–●●, 2008.

[25] L. Xie, L. Kennedy, S. Chang, A. Divakaran, H. Sun, and C. Lin, “Layered dynamic mixture model for pattern discovery in asynchronous multi-modal streams,” Peterson and MIT X Consortium. Athena Widgct Set C Language Itcfacc X Window System. MIT X Consortium, pp.●●–●●, 2005.

[26] H. Yanai and S. Puntanen, “Partial canonical correlations associated with the inverse and some generalized inverses of a partitioned dispersion matrix,” Statistical Sciences and Data Analysis: Proceedings of the Third Pacific Area Statistical Conference, eds. by K. Matsushita, M.L. Puri, and T. Hayakawa, pp.253–264, VSP International Science Publishers, 2003.

[27] D.R. Hardoon, S. Szedmak, O. Szedmak, and J. Shawe-taylor, “Canonical correlation analysis; an overview with application to learning methods,” Technical report ●●, , 2007.

[28] 中山英樹，原田達也，國吉康夫，“大規模 web 画像のための画像アノテーション・リトリーバル手法，” 電子情報通信学会論文誌. D, 情報・システム，vol.93，no.8，pp.1267–1280，2010.

[29] L.E. Baum and T. Petrie, “Statistical inference for probabilistic functions of finite state Markov chains,” The Annals of Mathematical Statistics, vol.37, no.6, pp.1554–1563,

1966.
[30] A.J. Viterbi, “Error bounds for convolutional codes and an asymptotically optimal decoding algorithm,” IEEE Transactions on Information Theory, vol.13, pp.260–269, 1967.
[31] Y.-G. Jiang, C.-W. Ngo, and J. Yang, “Towards optimal bag-of-features for object categorization and semantic video retrieval,” ACM International Conference on Image and Video Retrieval (CIVR’07), pp.●●–●●, 2007.
[32] M. Naphade, J.R. Smith, J. Tesic, S.F. Chang, W. Hsu, L. Kennedy, A. Hauptmann, and J. Curtis, “Large-scale concept ontology for multimedia,” ●●, vol.13, pp.86–91, 2006.
[33] M.R. Naphade, L. Kennedy, J.R. Kender, J.R.S. S-F Chang, P. Over, and A. Hauptmann, “A light scale concept ontology for multimedia understanding for trecvid 2005,” IBM Research Technical Report, pp.●●–●●, 2005.