論　文

映像の顕著性変動モデルを用いた視聴者の集中状態推定

米谷　竜[†a)]　川嶋　宏彰[†b)]　加藤　丈和[†c)]　松山　隆司[†d)]

Modeling Video Saliency Dynamics for Viewer State Estimation

Ryo YONETANI[†a)], Hiroaki KAWASHIMA[†b)], Takekazu KATO[†c)], and Takashi MATSUYAMA[†d)]

あらまし　本稿は，米谷竜, 川嶋宏彰, 加藤丈和, 松山隆司: ”映像の顕著性変動モデルを用いた視聴者の集中状態推定”, 電子情報通信学会論文誌, J96-D(8), pp.1675-1687, Aug. 2013 (http://search.ieice.org/bin/summary.php?id=j96-d_8_1675&category=D&year=2013&lang=J&abst) の著者バージョンです．図 1, 4, 7 を含め本研究で利用している映像はパナソニック株式会社の協力のもと提供されたものです．映像視聴における人間の視線運動には，その時々における視聴者の状態および映像中のシーンの特性が反映される．本研究では，映像のシーン特性と視線情報の関係性に基づいて視聴者の集中状態を推定することを目的とし，映像中の顕著領域が織りなす変動（顕著性変動）を線形システムにより表現する scene-based saliency dynamics model (SSDM) を提案する．提案手法では映像のシーン表現として，物体カテゴリといった多様性を持つ意味的情報ではなく，いくつかの典型的な顕著性変動パターンを導入する．これにより，映像の多様性を許容しつつシーン（変動パターン）の特性を考慮した視線解析を行うことが可能となる．本論文では，SSDM およびそのモデル推定法を提案するとともに，集中状態推定において SSDM が有効に働くことを示す．

キーワード　顕著性変動，saliency map，視線解析，集中状態推定

1. まえがき

人間の視線運動は，さまざまな高次認知処理を反映した複雑な現象である．視線運動の解析は視覚心理分野などで古くから取り組まれており，同一画像に向けられた視線運動が複数人あるいは個人の複数回の試行において類似することや，同一画像であっても人間の意図によって視線運動が異なることが実験的に明らかになっている [1]．これらの知見は，視線運動を統計的に解析することで人間に関するさまざまな状態が推定できることを示唆しており，近年における視線計測技術の発達にともない，視線情報の統計的学習に基づく人間の状態推定手法が各種提案されている [2], [3]．本研究では，人間が一般映像を視聴する状況をとりあげ，その際の視線情報から視聴者の状態，特に映像に対する集中状態を推定する問題に取り組む(注1)．提案手法は与えられた映像と視線データのペアに対して高集中，低集中といった集中状態ラベルをオフラインで与えることが可能であり，視聴者のプロファイリング，映像コンテンツへのメタデータ付与といった工学的応用に繋がることが期待される．

視線運動は人間の状態のみならず，視線の向けられたシーン，すなわち “人間がどのようなものを見ているのか” によっても多様に変化する．したがって，映像視聴中の視線情報を解析する上では，“映像中にどのようなものが映っているか” という映像のシーン特性を理解することが重要となる．シーンの認識・理解はコンピュータビジョン・パターン認識分野における中心的な課題であり，特定物体検出，認識では既に実用レベルの技術が提案されている [5], [6]．また，物体の持つ動きの情報は映像理解において重要な要素であり，多くの解析手法が提案されている [7]〜[14]．

しかしながら，一般物体認識の問題において取り上

† 京都大学 大学院情報学研究科
Graduate School of Informatics, Kyoto University
a) E-mail: yonetani@vision.kuee.kyoto-u.ac.jp
b) E-mail: kawashima@i.kyoto-u.ac.jp
c) E-mail: tkato@vision.kuee.kyoto-u.ac.jp
d) E-mail: tm@i.kyoto-u.ac.jp

(注1)：本論文は米谷ら，“映像の顕著性変動モデルを用いた視聴者の集中状態推定” [4] の拡張版である．

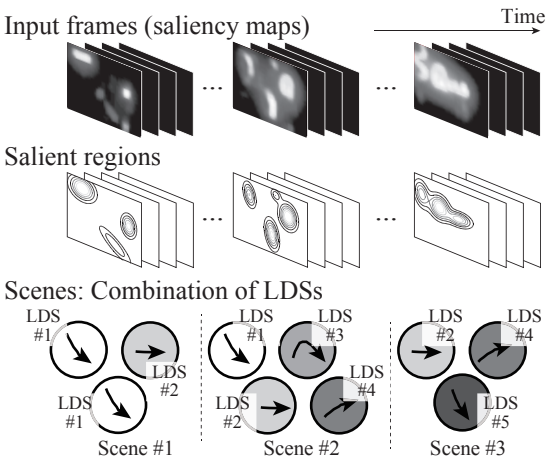


図 1 Scene-based saliency dynamics model.
Fig. 1 Scene-based saliency dynamics model.

げられるように [15]，本研究で扱うような一般映像には多様性がある．すなわち，映像のシーン中には膨大な種類の物体が存在し，かつ物体の配置や姿勢，その見えは時間とともに大きく変化しうる．したがって，一般映像のシーン特性を考慮しつつ視線運動を解析する上では，このように多様性を含む動的変化をモデル化し，視線情報と関連づける枠組みが必要となる．

これに対して我々は，視線情報に関連する特性として顕著性（saliency）に着目し，そのダイナミクスをモデル化して視線解析に用いる枠組みを提案している [16], [17]．映像には人間の視覚的注意を引きつける顕著領域が複数存在し，それらの位置や形状，顕著度は時間とともに変化する．このような複数領域によって織りなされる変動（**顕著性変動**）を映像から抽出し，映像のシーン表現としてその変動パターンをいくつかに分類して用いることで，映像の意味的な多様性を許容しつつシーン（変動パターン）ごとにその特性を考慮した視線解析が可能になる．

上述の枠組みのもとで，本研究では顕著領域の変動パターンを線形システムにより表現する *scene-based saliency dynamics model (SSDM)* を提案する．顕著領域の変動パターンはそれぞれ独立のダイナミクスに従い，かつフレーム中の領域数は時間とともに変化する．SSDM では顕著性変動の持つこのような特性を考慮し，顕著領域数および各領域のダイナミクスに基づいてシーンを記述する．具体的には，映像をシーンに分割し，各シーンの顕著性変動を線形システムの組み合わせにより記述することで，映像全体の顕著性変動を線形システム集合の切り替わりによって表現する（図 1）．これは先行研究 [17] の考え方を発展させたものであり，シーン中の領域数や線形システムのパラメタ同定において顕著性変動パターンに対する線形システムの当てはまりの良さを考慮することで，モデル推定の高精度化および顕著性変動モデルを用いた集中状態推定の性能向上を図っている．

つづく 2. 節では，関連研究と本研究の位置づけを概観する．3. 節，4. 節では，SSDM およびそのモデル推定法について具体的に説明する．5. 節では，SSDM を用いた視聴者の集中状態推定法を提案する．6. 節では，TV コマーシャル映像を用いた評価実験を通して，集中状態推定における SSDM の有効性を検討する．

2. 関 連 研 究

2.1 視線解析と人間の状態推定

視線運動と人間の状態（与えられたタスクや認知心理的な状態など）の分析は視覚心理分野において広く取り組まれているトピックであり，固視（fixation）や跳躍運動（saccade）の頻度や持続長といった視線特徴を統計的に解析するアプローチがしばしばとられる [2], [3]．また，同分野では一般映像視聴における視線運動についても研究されており，たとえば映画や自然映像を閲覧する際に複数人の視線運動がどの程度似通うかという問題が扱われている [18], [19]．その一方で，視対象の持つ特性を考慮しつつ視線解析を行うという枠組みは，主にヒューマンインタフェース分野で注目されており，インタラクティブシステム設計 [20]〜[23] や運転支援 [24]〜[27]，医療支援 [28], [29] などを目的としていくつかの手法が提案されている．これらの研究では考慮すべきシーンが比較的限定されており，その特性は手動，あるいは簡単な物体検出技術によって与えられることが多い．

これに対して本研究や [16], [17] では，一般映像のダイナミクスをモデル化・自動抽出し，視線運動の解析および人間の内的状態（ここでは集中状態）推定に用いることを目指す．具体的には，視線情報に関連する映像ダイナミクスとして顕著性変動に着目し，その変動パターンをシーン表現として用いる．そして，シーン特性を考慮するような視線特徴を設計し，内的状態との統計的な関係性をあらかじめ学習しておく．これにより，新たに得られた映像・視線のペアから内的状態の推定を行うことが可能となる．

このような顕著性変動モデルを用いた視線解析の枠

組みでは，以下の 2 点が問題となる．

(A) 顕著性変動をどのようにモデル化するか（シーンをどのように記述するか）

(B) 与えられたシーン特性を考慮するような視線特徴をどのように設計するか

[16] はこのうち (B) を中心に扱っており，シーンをフレーム中の顕著領域数（単数，複数）およびその動き（静止，移動）に基づいて記述し，シーン特性に応じて異なる視線特徴をトップダウンに組み合わせるアプローチを提案している．これに対して本研究は [17] と同様に主に (A) に着目し，より詳細なシーン記述や視線特徴の設計を目的として，顕著領域数およびそれら領域の位置や形状，顕著度が持つダイナミクスのモデル化を行う．ただし，本研究はモデル推定に関して [17] とは異なる手法をとる（詳細は 4. 1 節）．

2. 2 映像ダイナミクスのモデル化

SSDM は，線形システム集合の切り替わりによってフレーム中の顕著領域数および各領域の変動パターンの変化を表現するモデルである．線形システム集合を用いて対象を表現するモデルとして，bag of dynamical systems [14] がある．これは自然言語処理や物体認識においてしばしば利用される bag of words のアナロジーであり，映像から多数の時空間パッチを抽出し，そこに現れる変動パターンを有限個の線形システムによりモデル化するものである．SSDM は対象の順序関係や位置情報を捨象する点は上述のモデルと同様であるが，対象の従うダイナミクスが時間とともに変化する状況を陽に表現する点に違いがある(注2)．

一方，単一の主体による変動パターンを線形システムの切り替わりを用いて表現するモデルとして switching linear dynamical system (SLDS) があり，人間の複雑な動きをモデル化する際にしばしば用いられる [7]〜[9]．これに対して SSDM は時間とともに顕著領域数が変化する状況をモデル化しており，ダイナミクスおよびその主体数（モデルの状態空間）自体が変化する状況を表現できる点で SLDS とは大きく異なる．

複数主体の切り替わりを含むダイナミクスを扱ったものとして，複数対象の追跡を目的とした文献 [12] がある．[12] では主体数の増減をベルヌイ試行によりモデル化し，各主体のダイナミクスを混合ディリクレ過程に基づいて同定するとともに，各ダイナミクスのパラメタをパーティクルフィルタにより逐次更新する．これに対して SSDM は映像のバッチ解析を前提としており，ダイナミクスおよびその主体数は映像全体に対するモデルの効率性に基づいて推定される（詳細は 4. 節）．[12] のような逐次的手法は対象追跡といったリアルタイム性が要求される処理に優れ，SSDM のようなバッチ的手法はシーン分類といった効率的表現が要求される処理に優れている．

3. SSDM を用いた顕著性変動のモデル化

3. 1 映像の顕著性変動

映像中には，人間の視覚的注意を引きつける顕著領域が存在する．視覚的注意に関しては様々な計算モデルが提案されているが，本研究では顕著領域の抽出にあたって，文献 [30] などで提案されている顕著性マップ（saliency map）を用いる．[30] では，入力映像から明度，色差，エッジ方向といった基礎的画像特徴量のコントラストを複数スケールで抽出，統合することで，各ピクセルに対して顕著度（視覚的注意を引きつける度合い）の与えられた顕著性マップを獲得する．以下では，フレーム t における入力画像より計算された顕著性マップを $s_t : \mathbb{N}^2 \to \mathbb{R}^+$ と表記する(注3)．

映像中にはしばしば複数の物体が映される．したがって，顕著性マップにも複数の顕著領域が含まれる．これらの顕著領域は，それぞれ位置や形状，顕著度の分布といった特徴が時間とともに変化する．さらに，変動パターンが従うダイナミクス，そして領域数自体もまた時間とともに変化する．複数の顕著領域が織りなすこのような変動を，本研究では顕著性変動と呼ぶ．

3. 2 Scene-based saliency dynamics model

顕著性変動のモデル化では，フレーム中の顕著領域数および領域の従うダイナミクスが時間とともに変化する状況を，シーンとしてどのように捉えるかが問題となる．本研究では，ある時区間における各領域の変動パターンをそれぞれ単一の線形システムにより表現し，複数領域の変動パターンに対する線形システム集合をシーンして定義する．線形システムを利用することで，一定時区間にわたる複数パラメタ（たとえば領域の位置，形状）の変化パターンを単一の状態として記述できるようになる．また，線形システムのパラメタはデータから学習可能であり，後段の視線解析や集

(注2)：提案モデルの表現対象はシーン中の顕著領域であり，対象の認識を目的とした一般の bag-of-words 表現と比較して，抽出される word 数やその種類が非常に少ない点にも注意されたい．

(注3)：顕著性マップは [31] により計算した．画像特徴には明度，色差，エッジ方向，フレーム差分に基づく動き情報を用いた．

中状態推定において，与えられた映像を表現するために十分な線形システムのみを用いることができる．

SSDM の説明にあたって，まず映像が K 個の時区間からなる系列 $\mathcal{I} = (I_1, \ldots, I_K)$ に分割されていることを仮定する（具体的なシーン分割法は 4.3 節で提案する）．時区間 $I_k = [i_{k1}, i_{k2}]$ における顕著性マップ $\{s_t \mid t \in I_k\}$ はそれぞれ C_k 個の顕著領域を含み，$c \in \{1, \ldots, C_k\}$ 番目の顕著領域が時区間 I_k において織りなす変動パターンが $\Theta^{(c,k)} = \left(\boldsymbol{\theta}^{(c)}_{i_{k1}}, \ldots, \boldsymbol{\theta}^{(c)}_{i_{k2}}\right) \in \mathbb{R}^{J \times (i_{k2} - i_{k1} + 1)}$ とベクトル系列で表されるものとする（$\boldsymbol{\theta}^{(c)}_t \in \mathbb{R}^J$ はフレーム t における c 番目の顕著領域の特徴）．すると，時区間 I_k における顕著性変動は，変動パターン集合 $\boldsymbol{\Theta}_k = \{\Theta^{(1,k)}, \ldots, \Theta^{(C_k,k)}\}$ によって表現できる．

SSDM ではこの変動パターン集合 $\boldsymbol{\Theta}_k$ を用いてシーン $\boldsymbol{w}_k$ を記述する．まず，変動パターンを表現するためのコードブックとして，P 個の線形システムからなる集合 $\mathcal{D} = \{D_1, \ldots, D_P\}$ を考える．すなわち $\boldsymbol{\Theta}_k$ の各要素を $\mathcal{D}$ のうちいずれかの線形システムによって表現する．そして $\boldsymbol{w}_k$ を，$D_p \in \mathcal{D}$ の同定された変動パターンの数を要素 w_{pk} として持つ P 次元ベクトル

$$\boldsymbol{w}_k = (w_{1k}, \ldots, w_{Pk})^{\mathrm{T}} \tag{1}$$

として定義する（したがって $\sum_p w_{pk} = C_k$）．これにより，K 個の時区間よりなる映像全体の顕著性変動が $\mathcal{W} = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_K)$ としてモデル化される．

コードブックの要素である線形システム D_p は，式 (2) に示す 1 次の多変量自己回帰モデルとして与える．

$$\boldsymbol{z}_t = M^{(p)} \boldsymbol{z}_{t-1} + \boldsymbol{b}^{(p)} + \boldsymbol{v}^{(p)}_t. \tag{2}$$

$\boldsymbol{z}_t$ は時刻 t における領域の状態ベクトル（すなわち $\boldsymbol{z}_t = \boldsymbol{\theta}^{(c)}_t$），$M^{(p)}$ は遷移行列，$\boldsymbol{b}^{(p)}$ はバイアス，$\boldsymbol{v}^{(p)}_t$ はガウス分布 $\mathcal{N}(0, Q^{(p)})$ によってモデル化されるノイズであり，D_P は $M^{(p)}, \boldsymbol{b}^{(p)}, Q^{(p)}$ をパラメタに持つ．

3.3 顕著領域のパラメタ表現

提案モデルの導入にあたって，顕著領域のパラメタ表現 $\boldsymbol{\theta}^{(c)}_t$ を定義する必要がある（c は顕著領域に対して与えられた ID，t はフレーム ID）．対象の形状ダイナミクスの表現手法としては Snakes [32] やレベルセット法 [33] が，またテクスチャダイナミクスの表現手法として dynamic textures [10], [14] などが挙げられるが，これらは領域の位置・形状およびテクスチャ（顕著度分布）のダイナミクスを同時に扱うことができない．一方，両者を結びつける枠組みとして文献 [11] や active appearance model (AAM) [34] があるが，前者は複数領域のダイナミクスを表現するものではない．また AAM は，モデル学習時における手動での特徴点選択にその精度が大きく依存するという問題がある．

そこで本研究では，フレーム中の複数顕著領域を混合正規分布（Gaussian mixture model; GMM）によってモデル化する．すなわち，各顕著領域を単一のガウス分布によって表現する．GMM によるモデル化は顕著領域の詳細な形状や顕著度分布を表現するものではないが，複数領域の位置，おおまかな形状および顕著度の大きさがそれぞれガウス分布の平均，分散，混合比として，一つのモデルにより表現可能になる．

GMM の推定は以下の手順で行われる．まず，入力となる顕著性マップ s_t を多数サンプルにより近似する．そして，GMM のパラメタを expectation-maximization（EM）アルゴリズムによって推定する．EM アルゴリズムは局所最適化の手法であり，初期値依存性が強い．また，顕著領域の変動パターンを線形システムでモデル化するにあたって，領域のパラメタは連続的に変化することが望ましい．そこで本研究では，ある時刻における GMM パラメタの推定結果を，次時刻の EM アルゴリズムの初期値として与える．ただし，シーン開始点では，EM アルゴリズムの初期値はランダムに与えるものとする．推定から得られた c 番目のガウス分布の平均，分散，混合比を，以下ではそれぞれ $\boldsymbol{\mu}^{(c)}_t, \Sigma^{(c)}_t, \phi^{(c)}_t$ と表す．

顕著性マップ系列を GMM 系列として表現することで，時区間 $I_k = [i_{k1}, i_{k2}]$ における C_k 個の領域変動パターンが，時空間的に最近傍となるガウス分布を追跡することにより得られるようになる．いま，シーン開始点において c 番目のガウス分布を追跡することで得られた平均，分散，混合比の時系列パターン $(\boldsymbol{\mu}^{(c)}_{i_{k1}}, \ldots, \boldsymbol{\mu}^{(c)}_{i_{k2}}), (\boldsymbol{\sigma}^{(c)}_{i_{k1}}, \ldots, \boldsymbol{\sigma}^{(c)}_{i_{k2}}), (\phi^{(c)}_{i_{k1}}, \ldots, \phi^{(c)}_{i_{k2}})$ を考える（$\boldsymbol{\sigma}^{(c)}_t \in \mathbb{R}^3$ は $\Sigma^{(c)}_t$ における分散，共分散成分）．ここでは，顕著領域の相対的な変動パターンに着目するため，また次節に述べるモデル推定の簡単化のため，各特徴の時系列パターンを標準化した後，主成分分析を適用し，第 1 主成分の変動パターン $(\bar{\mu}^{(c)}_{i_{k1}}, \ldots, \bar{\mu}^{(c)}_{i_{k2}}), (\bar{\sigma}^{(c)}_{i_{k1}}, \ldots, \bar{\sigma}^{(c)}_{i_{k2}}), (\bar{\phi}^{(c)}_{i_{k1}}, \ldots, \bar{\phi}^{(c)}_{i_{k2}})$ を領域の特徴として用いる（$\bar{\phi}^{(c)}_t$ は標準化のみ行う）．そして，変動パターン $\Theta^{(c,k)}$ 中の要素 $\boldsymbol{\theta}^{(c)}_t$ を $\boldsymbol{\theta}^{(c)}_t = (\bar{\mu}^{(c)}_t, \bar{\sigma}^{(c)}_t, \bar{\phi}^{(c)}_t)^{\mathrm{T}}$ として定義する．

4. SSDM のモデル推定

4.1 問題設定

本研究で提案する SSDM は，シーン系列 $\mathcal{W} = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_K)$ により映像全体の顕著性変動を表現するものであり，同モデルの推定では映像のシーン分割推定（時区間系列 $\mathcal{I} = (I_1, \ldots, I_K)$ の推定）およびシーン表現のためのコードブック $\mathcal{D} = \{D_1, \ldots, D_P\}$ の生成が必要となる．

3.2 節で述べた通り，シーン $\boldsymbol{w}_k$ は $\mathcal{D}$ を用いて顕著領域数および領域のダイナミクスを表現する．そのためシーン分割 $\mathcal{I}$ は，時区間 I_k 中の変動パターン $\boldsymbol{\Theta}_k = \{\Theta^{(1,k)}, \ldots, \Theta^{(C_k,k)}\}$ に当てはまりの良い（予測誤差の小さい）線形システム $D_p \in \mathcal{D}$ が同定されるよう与えられることが望ましい．その一方で，変動パターンに対して線形システム D_p を同定し，当てはまりの良さを評価するためには，シーン分割 $\mathcal{I}$ が既知である必要がある．先行研究 [17] のモデル推定では，与えられた映像全てをバッチ解析することで線形システムのパラメタ推定を行うものの，顕著領域数は貪欲法（greedy algorithm）に基づいて直前フレームの情報のみを考慮して推定するため，変動パターンに対する線形システムの当てはまりの良さを考慮しつつシーンを分割することは難しい．

これに対して本研究では，シーン時区間候補を多数生成し，各候補における線形システム同定コストに基づいて適切なシーン分割を選択する．具体的にはまず，顕著性マップのフレーム間差分系列を多重解像度表現することでシーン分割の階層構造を生成し（図 2 (2-1)），そこから得られる複数の時区間候補においてそれぞれ GMM のフィッティングおよび変動パターンに対する線形システム同定を行う（図 2 (1)）．そして，隣接時区間候補により定義されるシーン分割点を，各時区間における線形システム同定コストに基づいて評価することで，顕著性変動のシーン分割を推定する（図 2 (2-2)）．結果として，線形システムの当てはまりの良さに基づいてシーン分割が行われることになる．

シーン分割を行うことで，同時に各領域の変動パターンに対して線形システムが同定される．これらをクラスタリングすることにより，コードブック $\mathcal{D} = \{D_1, \ldots, D_P\}$ が生成できる．そして，$\mathcal{D}$ に基づいて各領域の線形システムに割り当てられたクラスタ ID を $\boldsymbol{w}_k = (w_{k1}, \ldots, w_{pk})^{\mathrm{T}}$ の対応要素に投票することで，顕著性変動がシーン系列 $\mathcal{W} = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_K)$ として表現されることになる．

つづく 4.2 節では，与えられたシーン時区間における線形システム集合の同定法および同定コストの算出法を提案する．そして 4.3 節では，線形システム同定コストを用いてシーン分割を行う手法を述べる．4.4 節ではコードブック生成のための線形システムクラスタリング法について述べる．

4.2 線形システム集合の同定

時区間 $I_k = [i_{k1}, i_{k2}]$ における変動パターン集合に対する線形システム集合の同定は，GMM のコンポーネント数 C_k の推定および各変動パターンに対する線形システム同定によって行われる（図 2 (1)）．

コンポーネント数の決定（図 2 (1-1)）

まず，時区間 I_k 中の顕著性マップ系列 $(s_{i_{k1}}, \ldots, s_{i_{k2}})$ に対して，あらかじめ取りうるコンポーネント数の範囲 $\{\Gamma_{\min}, \Gamma_{\max}\}$ を定める．そして，3.3 節で述べた方法により各フレームの顕著性マップ s_t についてコンポーネント数 $\Gamma_{\min}, \ldots, \Gamma_{\max}$ の GMM を当てはめる．この際，コンポーネント数 Γ の GMM によって s_t をモデル化する際のコスト（GMM 化コスト）$e_t^{(\Gamma)}$ を，赤池情報量規準（Akaike information criterion; AIC）によって定める．すると，顕著性マップ系列 $(s_{i_{k1}}, \ldots, s_{i_{k2}})$ をコンポーネント数 Γ の GMM の系列としてモデル化する際の GMM 化コスト $E_k^{(\Gamma)}$ は，各フレームにおける GMM 化コストの和 $E_k^{(\Gamma)} = \sum_{t=i_{k1}}^{i_{k2}} e_t^{(\Gamma)}$ として表現できる．時区間 I_k におけるコンポーネント数 C_k は，$E_k^{(\Gamma)}$ を最小にする $\Gamma = \hat{\Gamma}_k$ として推定される．

線形システムの同定（図 2 (1-2)）

$\Theta^{(\gamma,k)}$ に対して同定される線形システムを $D^{(\gamma,k)}$ と表記する（$\gamma \in \{1, \ldots, \hat{\Gamma}_k\}$）．本研究では線形システムとして式 (2) に示すような 1 次の多変量自己回帰モデルを仮定しているため，$\boldsymbol{\theta}_{t-1}$ から $\boldsymbol{\theta}_t$ を予測する誤差が小さくなるように線形システムの遷移行列 $M^{(\gamma,k)}$ およびバイアス $\boldsymbol{b}^{(\gamma,k)}$ を推定すればよい．ただし，$\Theta^{(\gamma,k)}$ によって表される顕著領域の位置および形状，顕著度は高い相関を持つことがあり，このような場合多重共線性の問題から線形システムのパラメタが正しく推定できない．そこでここでは，$\Lambda = [M^{(\gamma,k)} \mid \boldsymbol{b}^{(\gamma,k)}]$ に関する正則化項 $||\Lambda||_{\mathrm{F}}^2$ を加えたリッジ回帰を最小二乗法で解くことにより，パラメタ推定を行う（$||\cdot||_{\mathrm{F}}$ は行列のフロベニウスノルム）．このとき，正則化パラメタは $||\Lambda||_{\mathrm{F}}$ が一定閾値以下になるように設定する．パラメタ Λ が推定されると，初期値 $\boldsymbol{\theta}_{i_{k1}}^{(\gamma)}$ を与えることで変動パターンの生成が可能

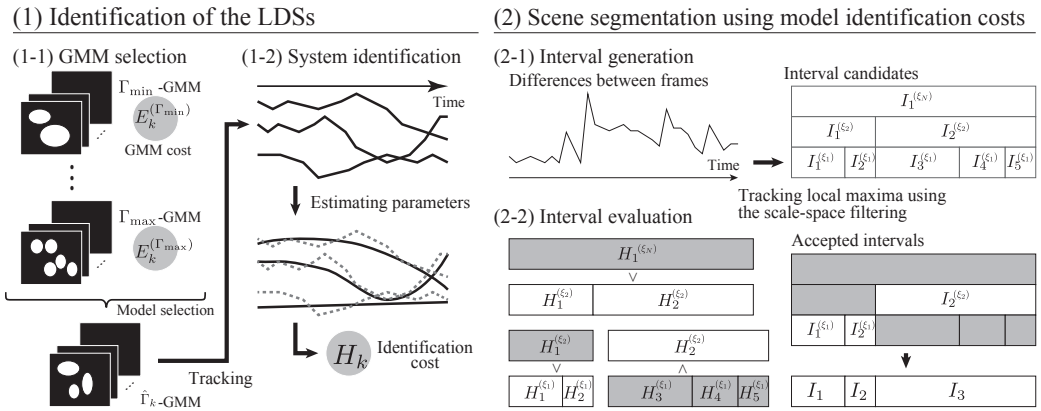


図 2　SSDM のモデル推定.
Fig. 2　Model estimation scheme for the SSDM.

になる．これを用い，$\Theta^{(\gamma,k)}$ から推定パラメタを用いて変動パターンを生成し，$\Theta^{(\gamma,k)}$ との誤差分布をガウス分布によってモデル化することにより，誤差共分散 $Q^{(\gamma,k)}$ を得る．そして，$\mathcal{N}(0, Q^{(\gamma,k)})$ に基づいて生成パターンに対する $\boldsymbol{\theta}_t$ の誤差を評価することで，推定パラメタに関する AIC $h^{(\gamma,k)}$ が計算できる．

こうして，時区間 I_k における $\hat{\Gamma}_k$ 個の変動パターンについて，線形システム集合 $\{D^{(1,k)}, \ldots, D^{(\hat{\Gamma}_k,k)}\}$ が同定される．I_k における線形システム同定コスト H_k は，全パターンのうち線形システムの当てはまりが最も悪いものを考慮し，得られた AIC の最大値（すなわち $H_k = \max_\gamma h^{(\gamma,k)}$）を与える．

4.3　多重解像度解析に基づく映像のシーン分割

シーン候補の生成（図 2 (2-1)）

映像の分割（ショット分割，カット検出）は映像解析においてしばしば扱われる問題であり，隣接フレームの差分や類似フレームのクラスタリングを用いたアプローチがある（たとえば [35]）．これに対して本研究のシーン分割は，各時区間における変動パターンが線形システムによって精度よくモデル化されるかを基準に行われるものであり，上記の分割手法とは異なるアプローチを導入する必要がある．

提案手法におけるシーン分割の基本的な考え方は，多重解像度解析の枠組みに基づきシーン時区間候補を複数生成しておき，各時区間候補における線形システム同定コストに基づいて分割点の評価を行うことでシーン分割を得るというものである．まず，隣接フレーム s_{t-1}, s_t 間の差分 $f_t \in \mathbb{R}^+$ を各ピクセル間差分の二乗和として定義する．そして，分割点検出のための入力として，フレーム間差分系列 $\boldsymbol{f} = (f_1, \ldots, f_T)$ を考える．系列 $\boldsymbol{f}$ に対して平滑化スケール $\{\xi_1, \ldots, \xi_N\}$ $(\xi_{n-1} < \xi_n)$ のガウス関数を畳み込むことによって，系列の多重解像度表現ができる（$*$ は畳み込み演算を表す）．

$$\boldsymbol{f}^{(\xi_n)} = \boldsymbol{f} * g^{(\xi_n)},\ g^{(\xi_n)} = \mathcal{N}(0, \xi_n) \qquad (3)$$

このとき，波形集合 $\{\boldsymbol{f}^{(\xi_1)}, \ldots, \boldsymbol{f}^{(\xi_n)}\}$ においてスケールを変化させながら変曲点を追跡すると，ガウス関数の持つ因果性（causality）により，スケールを小さくしていくにつれて新たな変曲点が現れるような階層構造が形成される．議論の簡単化のため，ここでは $\{\xi_1, \ldots, \xi_N\}$ を，すべてのスケール変化 $\xi_n \to \xi_{n-1}$ において新たな変曲点が現れるように設定する．また，最高スケール ξ_N は変曲点を持たないスケールとする．

以下では，各スケールの変曲点集合について，各変曲点の追跡により得られるスケール ξ_1 での変曲点をそのスケールにおけるシーン分割候補点として想定し，候補点に定義される複数分節をそれぞれシーン時区間候補として扱う．スケール ξ_n において形成された時区間候補を $\hat{\mathcal{I}}^{(\xi_n)} = (\hat{I}_1^{(\xi_n)}, \ldots, \hat{I}_{K_{\xi_n}}^{(\xi_n)})$ と表記する．4.2 節で述べた手法によって，$\hat{I}_k^{(\xi_n)}$ の変動パターンに対して線形システム集合が同定でき，その際の同定コスト $H_k^{(\xi_n)}$ が当該時区間に対して与えられる．

シーン分割の推定（図 2 (2-2)）

すべてのシーン時区間候補に対して線形システム同定コストを与えることで，シーン分割点の評価が可能になる．スケール ξ_n における時区間候補 $\hat{I}_k^{(\xi_n)}$ と同一区間に定義されるような，スケール ξ_{n-1} における $\hat{\mathcal{I}}^{(\xi_{n-1})}$ の部分系列 $\hat{\mathcal{I}}^{(\xi_{n-1})}|_{(j,j+l)} = (I_j^{(\xi_{n-1})}, \ldots, I_{j+l}^{(\xi_{n-1})})$ を

考える．このとき，当該区間を時区間（系列）$\hat{I}_k^{(\xi_n)}$，$\hat{\mathcal{I}}^{(\xi_{n-1})}|_{(j,j+l)}$ のいずれで表すか（当該時区間において分節化を行うか）を，線形システム同定コストによって判断する．すなわち，もし $H_k^{(\xi_n)} > \sum_{j'=j}^{j+l} H_{j'}^{(\xi_{n-1})}$ ならば分節化を行う．このような評価を，最高スケール ξ_N における時区間 $\hat{\mathcal{I}}^{(\xi_N)}$ から再帰的に行うことで，変動パターンを線形システムによって表現するために適切なシーン分割を得ることができる．

4.4 線形システムの階層クラスタリング

前節までの手続きによって，映像が時区間系列 $\mathcal{I} = (I_1, \ldots, I_K)$ に分節化され，かつ時区間 I_k における顕著性変動が $\hat{\Gamma}_k$ 個の要素よりなる線形システム集合 $\{D^{(1,k)}, \ldots, D^{(\hat{\Gamma}_k,k)}\}$ によってモデル化される．この段階では，すべての時区間について，各線形システムは独立に同定されている．これらの線形システムをクラスタリングすることで，P 個の線形システムからなるコードブック $\mathcal{D} = \{D_1 \ldots, D_P\}$ を生成することができる．そして，$\{D^{(1,k)}, \ldots, D^{(\hat{\Gamma}_k,k)}\}$ において D_p に分類されたシステム数を w_{pk} とすることで，変動パターンが式 (1) に定義したシーン $\boldsymbol{w}_k$ として表現できることになる．

映像（複数映像でも良い）中の顕著領域の変動パターンから，合計 P' 個の線形システム $\mathcal{D}' = \{D'_1 \ldots, D'_{P'}\}$ が同定されたとする．これらの線形システムに対してクラスタリング処理を行うことで，$P < P'$ 個の線形システムからなるコードブックを生成する．このとき，システム間の非類似度を何らかの方法で定義する必要がある．本研究では，各線形システムが変動パターンを生成できることに着目し，システムから生成された固定長パターンをシステムを表現する特徴ベクトルとして扱い，パターン同士の相関を非類似度として定める．これにより，直感的に動きの似通った線形システムが同一クラスタになることが期待できる．

線形システム $D'_{p'}$ から生成された固定長 l の変動パターン $\hat{\Theta}_{p'} \in \mathbb{R}^{3\times l}$ を考える．この変動パターンは顕著領域の位置，形状，顕著度（位置，形状はその主成分）よりなる 3 次元の時系列パターンである．$\hat{\Theta}_{p'}$ を行ベクトルに分解したものを $\hat{\Theta}_{p'} = ((\boldsymbol{\eta}_1^{(p')})^{\mathrm{T}}, (\boldsymbol{\eta}_2^{(p')})^{\mathrm{T}}, (\boldsymbol{\eta}_3^{(p')})^{\mathrm{T}})^{\mathrm{T}}$ と表す（$\boldsymbol{\eta}_j^{(p')}$ は長さ l の行ベクトル）．このとき，パターン $\hat{\Theta}_{p'}, \hat{\Theta}_{p''}$ の非類似度 $Z(\hat{\Theta}_{p'}, \hat{\Theta}_{p''})$ を式 (4) により定義する．

$$Z(\hat{\Theta}_{p'}, \hat{\Theta}_{p''}) = 1 - \sum_{j=1}^{3} \omega_j \mathrm{ZNCC}(\boldsymbol{\eta}_j^{(p')}, \boldsymbol{\eta}_j^{(p'')}) \quad (4)$$

ただし，ZNCC は正規化相互相関，ω_j $(\sum_j \omega_j = 1)$ は各パラメタに対する重みを表す．

この非類似度に従って線形システムの階層クラスタリングを行い，クラスタ数ごとにクラスタリングにかかるコスト（クラスタリングコスト）を計算することで，コードブックの線形システム数 P を決定する．ここでは，階層クラスタリング手法の一つである完全リンク法（complete linkage）に基づき，線形システム集合 $U, V \subset \mathcal{D}'$ 間の非類似度には，変動パターン $\Theta_u \in U, \Theta_v \in V$ 間の非類似度 $Z(\Theta_u, \Theta_v)$ の最大値を用いる．またクラスタリングコストは，同一クラスタの変動パターン集合から一つの線形システムを再同定する際の AIC の和として定める．

クラスタリングコストに基づく最終的な線形システム数 P の決定について，本研究では文献 [36], [37] に倣い，クラスタ数の減少にともなうクラスタリングコストの増加が急になる直前（表現の冗長性が十分小さくなった時点）のクラスタ数を P として手動で定める．コードブック $\mathcal{D} = \{D_1, \ldots, D_P\}$ は，各クラスタの変動パターン集合からそれぞれ一つの線形システムを再同定することによって得る．

5. 顕著性変動モデルを用いた集中状態推定

本節では，モデル化された顕著性変動に基づき，視線特徴量の統計的学習に基づいて映像視聴者の集中状態推定を行う手法について説明する．

5.1 映像視聴状況における集中状態

ディスプレイに映像が提示され，人間がそれを視聴する状況を想定する．ディスプレイ下には視線計測装置が設置され，視線運動（ディスプレイ上の注視点系列）を精度良く計測できるものとする．このとき，視聴者が映像に対してどれだけ注意を向けているかという注意の程度を集中の高さとして定義し，集中の高さは数段階に量子化できると仮定する．集中状態推定とは，映像および計測された視線運動に基づいて，視聴者の集中の高さを推定する問題である [16], [17].

5.2 視線特徴量の学習と集中状態推定

対象映像が，あらかじめ学習されたコードブック $\mathcal{D}$ を用いてシーン系列 $\mathcal{W} = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_K)$ として表現されているとする．[16] では，視線運動が映像のシーン特性に影響を受けることに着目し，シーンによって異なる視線特徴（たとえば，顕著領域の注視時間や跳躍運動の頻度）をトップダウンに設計している．しかしながら本研究において，シーンは線形システムの組

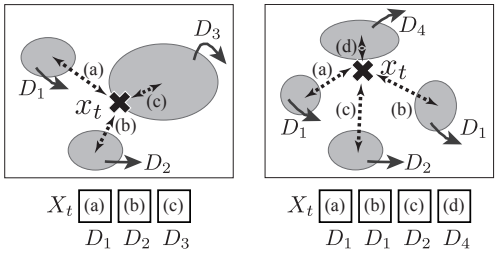


図 3　局所的顕著性変動の抽出.
Fig. 3　Local saliency dynamics.

み合わせによって記述されており，その種類数は [16] と比較してより多くなるため，同様のアプローチを取ることは難しい.

そこで本研究では，生成されたシーンの多様性を許容しつつその特性を反映する視線特徴として，注視点近傍にどのような顕著性変動パターンが存在するかという**局所的顕著性変動**の情報を用いる．具体的には，図 3 のように線形システムの同定された各顕著領域と注視点との位置関係を，局所的顕著性変動を表現する視線特徴として抽出する．いま，シーン $\boldsymbol{w}_k$ を構成する C_k 個の顕著領域はそれぞれガウス分布でモデル化されている．ここで，時刻 t における分布中心が $(\boldsymbol{\mu}_t^{(1)}, \ldots, \boldsymbol{\mu}_t^{(C_k)})$ と表され，$(\boldsymbol{\mu}_t^{(1)}, \ldots, \boldsymbol{\mu}_t^{(C_k)})$ は当該領域の変動パターンに同定された線形システムの ID $(1, \ldots, P)$ が小さな順に並びかえられているとする（ID が重複する場合 $\boldsymbol{\mu}_t^{(c)}$ はラスタ順に並ぶものとする）．このとき，時刻 $t \in I_k$ における注視点 $\boldsymbol{x}_t \in \mathbb{R}^2$ を用いて視線特徴 X_t を以下のように定義する.

$$X_t = (||\boldsymbol{x}_t - \boldsymbol{\mu}_t^{(1)}||_2, \ldots, ||\boldsymbol{x}_t - \boldsymbol{\mu}_t^{(C_k)}||_2) \quad (5)$$

X_t は全シーンについて式 (5) より計算できる点で，生成されたシーンの多様性を許容している.

このようにして計算された視線特徴 X_t は，集中状態 A，シーン $\boldsymbol{w}_k$ ごとに投票，確率密度の推定を経て $P(X = X_t \mid A, \boldsymbol{w} = \boldsymbol{w}_k)$ として統計的に学習される．これにより，新たな映像，視線データを獲得した際，集中状態の識別が可能になる．まずコードブック $\mathcal{D}$ を用いて映像をシーン系列 $\mathcal{W}' = (\boldsymbol{w}'_1, \ldots, \boldsymbol{w}'_{K'})$ に変換する．そして $t \in I'_k$ における視線特徴 X'_t を計算し，式 (6) により尤度に基づく集中状態の識別を行う.

$$\hat{A} = \arg\max_A P(X = X'_t \mid A, \boldsymbol{w} = \boldsymbol{w}'_k) \quad (6)$$

6. 実　　験

6.1　一般映像に対する SSDM の適用

SSDM を一般映像に適用した例を示す．本研究では, [16], [17] と同様の TV コマーシャル 12 種類（いずれも映像時間は 15 秒）を用いた．TV コマーシャルはもともと視聴者の視覚的注意を引きやすいデザインがなされており，明確な顕著領域が存在することが期待できる．4.2 節で述べた GMM のコンポーネント数の決定に関して，ここでは $\Gamma_{\min} = 1, \Gamma_{\max} = 8$ に設定した．また SSDM のモデル推定に関して，4.4 節のコードブック生成には 12 種類の映像すべてを用い，式 (4) の重み ω_j は $\omega_1 = \omega_2 = \omega_3 = 1/3$ に設定した.

図 4 に顕著性マップ系列および当てはめられた GMM，そこから得られる顕著性変動パターンを示す．本実験では，12 種類の映像より 197 個のシーン時区間，合計 782 個の変動パターンが得られた．またコンポーネント数 C_k は，いずれも $3 \leqq C_k \leqq 5$ となった．そこから線形システムのクラスタリングを行い，図 5 に示すクラスタリングコストの変化に基づいて $P = 8$ のコードブックを生成した．実験において生成されたコードブックおよび図 4 に示した変動パターンに対するモデル推定結果を図 6 に示す．生成された線形システムはいずれも，顕著領域の位置（mean），形状（covariance）および顕著度（weight）の単純な変化（たとえば領域の移動や拡大・縮小，顕著度の増減）の組み合わせを表現している．本実験では合計 142 種類のシーンが獲得された.

6.2　提案モデルを用いた集中状態推定

a）**実 験 設 定**

集中状態推定を通して提案手法の性能を確認する．映像コンテンツおよび視線データには文献 [16], [17] と同様のものを利用した．本データセットは先述の TV コマーシャル 12 種類より構成されており，10 名の被験者はそれぞれ以下の 2 条件で映像視聴を行っている.

条件 1　映像を視聴し，その後アンケートに回答する.
条件 2　映像視聴と同時に暗算タスクを遂行する.

条件 2 では集中のリソースが暗算タスクに割かれることから，映像視聴への集中が条件 1 と比較して低くなっていると見なす．すなわち，条件 1，2 がそれぞれ映像視聴に対する高集中，低集中状態に対応する．本実験ではこれらの 2 状態の識別を通して手法の性能を検証する（詳細は付録 1. を参照）.

推定精度の評価は，10 名 ×12 映像 ×2 条件 =240

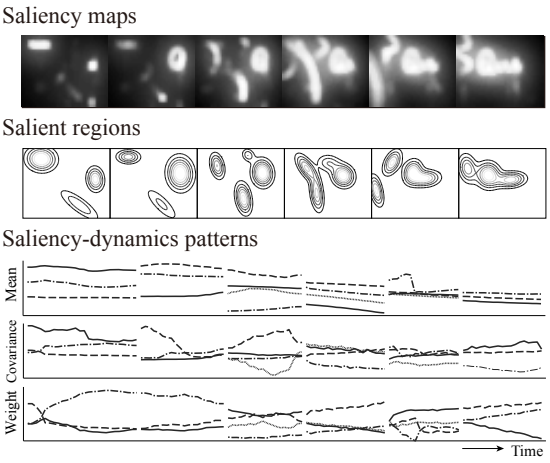


図 4　顕著性マップ系列より抽出された顕著性変動パターン．1 段目: 入力の顕著性マップ系列（各シーンの第 1 フレーム），2 段目: GMM の当てはめ結果，3〜5 段目: 顕著性変動パターンを表す．

Fig. 4　Example of saliency dynamics patterns. The 1st and 2nd rows depict selected saliency maps and the corresponding results of GMM fitting. The rest of the rows depicts the time-varying patterns of salient regions.

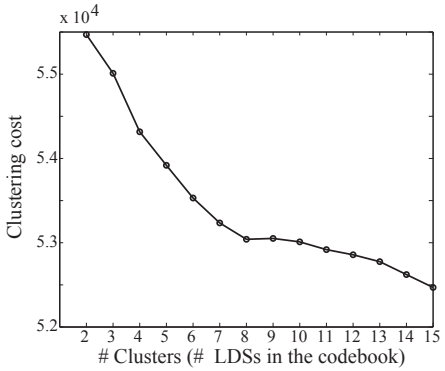


図 5　クラスタ数（コードブックの線形システム数）に対するクラスタリングコストの変化．

Fig. 5　Changes of clustering costs for the number of clusters (the number of linear dynamical systems in the codebook).

試行（各集中状態につき 120 試行）のデータに対して leave one out 法を用いることで行った．集中状態推定は，得られた視線特徴量分布に対してカーネル密度推定を行った後，単純ベイズ法により行った．カーネルのパラメタは学習データ内の交差検定により決定した．

b）**評価基準と比較手法**

本実験では集中状態の推定精度を通して，(1) 顕著領域の抽出において SSDM を用いること，(2) 視線特

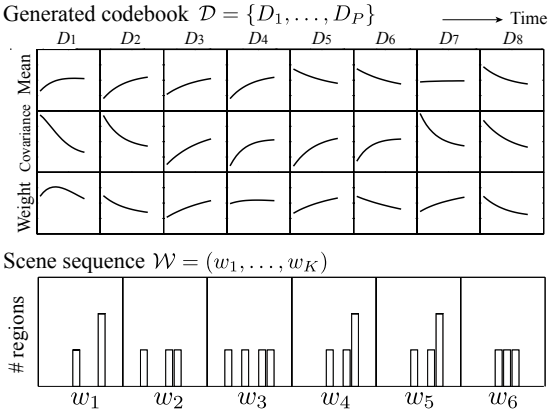


図 6　生成されたコードブック（上段）および図 4 に示した変動パターンに対するモデル推定結果（下段）．下段: 縦軸は各線形システムの同定された変動パターン数，横軸は線形システムの ID を表す．

Fig. 6　Generated codebook (the top row) and model estimation results for saliency-dynamics patterns in Fig. 4 (the bottom row).

徴の設計においてシーン特性（顕著性変動）を考慮すること，そして (3) 視線特徴の設計において SSDM を用いることの有効性を検証する．まず (1) に関して，ここでは視線特徴として顕著領域の注視時間（ある時間窓において領域内を注視する割合）を用い，[16] の顕著性変動モデルを利用して顕著領域を抽出した場合（**BM1**）および SSDM を利用する場合（**BM2**）を比較した．これらの手法では，視線特徴量はシーンの種類によらず集中状態ごとに学習した．[16] では各フレームの顕著性マップを二値化することで顕著領域を抽出しており，その際に領域の数や変動パターンといった情報は用いられない．これに対して SSDM では，領域の変動パターンに対する線形システムの当てはまりの良さに基づいて領域数とそのパラメタを推定している．なお SSDM において顕著領域は GMM によりモデル化されるため，視線特徴の抽出にあたっては GMM の事後確率分布を適当な閾値で二値化して用いた．

次に (2) に関して，**BM1** と [16] の提案手法（**PM1**）および **BM2** と本研究における提案手法（**PM2**）を比較評価した．また **PM1** と **PM2**，先行研究 [17] の提案手法（**PM2'**）を比較することで (3) を検証した．視線特徴の設計に関して，**PM1** は領域の注視時間や跳躍運動の頻度といった視線ダイナミクスの特性を考慮し，**PM2**，**PM2'** は注視点近傍の局所的顕著性変動という映像ダイナミクスの特性を考慮するという違いがある．また，4. 節に述べたように，**PM2** と

表 1　実験結果.
Table 1　Experimental results.

Method	BM1	BM2	PM1	**PM2**	PM2'
Accuracy (%)	53.8	65.0	78.2	**85.4**	80.6

PM2' はモデル推定法が異なる.

c）**結果および討議**

表 1 に推定精度を示す．上に挙げた項目 (1)，(2)，(3) いずれにおいても，シーン特性を考慮する場合や SSDM を用いる場合がより高精度であり，提案手法の有効性が確認された.

PM2 における顕著性変動のモデル化，学習された視線特徴量の一例を図 7 に示す．図中 2 列目では，局所的顕著性変動を構成する線形システムのうち集中状態推定に寄与したもの，すなわち集中状態によって注視される程度に差が見られた線形システムを示している．この結果から分かるように，どの線形システムが推定に寄与するかはシーンによって異なっており，視線特徴の設計において SSDM から得られるシーン特性を考慮することの有効性が確認できる.

また，学習された視線特徴量を可視化するため，集中状態をクラスとしたフィッシャー線形判別を行い，クラスの分離を最大にする射影軸を求めた．図中 3 列目では，求めた軸上に射影された視線特徴量分布を示している．このように，集中状態に関する視線特徴量分布の分離度はシーンによって異なる．これは，提案する枠組みがどの程度有効に働くかがシーンに依存することを示唆している.

なお **PM1**，**PM2** に共通して，学習セット中に存在しない（視線特徴量の学習されていない）シーンに対しては推定が行えないという限界が存在する．本実験では学習に際して全ての映像を用いたが，未知映像に対する集中状態推定を行う場合，視線特徴量の学習済みシーンの割合が重要となる．一つの方法として，推定精度が保証される範囲でコードブック $\mathcal{D}$ のサイズ P（線形システム数）を小さくすることで学習済シーンの割合を増加させるアプローチが考えられる．本実験では $P = 8$ に設定されており，学習済シーンは全体の 11.4%であった(注4)．これに対して，たとえば $P = 5$ の場合，推定精度は 81.1%と低下するものの，

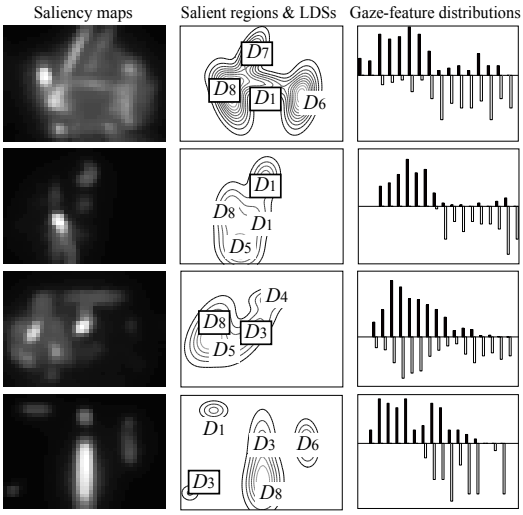


図 7　SSDM による顕著性変動のモデル化および学習された視線特徴量．左から順に顕著性マップ，モデル推定の結果（太枠は集中状態推定に寄与した線形システムを示す），学習された視線特徴量をフィッシャー線形判別に基づいて判別軸に射影した分布（上: 高集中，下: 低集中）を示す.

Fig. 7　Illustrative results. The 1st and 2nd columns depict saliency maps and model estimation results. LDSs with bold rectangles contribute to attentive-state estimation. The 3rd column shows the gaze-feature distributions after the Fisher's discriminant analysis (above: high attentive, bottom: low attentive).

映像から得られるシーンは 97 種類まで減少し，全体のうち 42.0%のシーンに対して視線特徴量の学習が可能になる．本研究では変動パターンに対する線形システム当てはめの誤差に基づいてコードブックを決定したが，集中状態推定の汎化性能向上を目的として，上述のように学習データに含まれるシーンの割合を考慮しつつコードブックを定めることも可能である.

なお，本研究で提案した SSDM は映像のシーン表現として顕著領域のダイナミクスのみを扱うものであり，領域が特定の物体（顔，文章など）を含んでいる，音声信号（ナレーションなど）が特定の領域に関連しているといった意味的特性は考慮していない．これに関して，たとえば [38] や [39] のように顕著性計算と特定物体検出を組み合わせて用いることで，顕著領域の意味的特性を考慮することができる．また，音声信号と映像ダイナミクスを関連づける手法 [40] や，音声信号自体の顕著性を考慮する手法 [41] を導入すること

(注4)：P 個の線形システムよりなるコードブック $\mathcal{D}$ を用いて C 個の変動パターンに同定されるシーンは，重複組合せにより最大 $\left(\begin{array}{c} P+C-1 \\ C \end{array} \right)$ 種類になりうる．本研究では $3 \leqq C \leqq 5$ と推定されており，シーンの種類数は $\sum_{C=3}^{5} \left(\begin{array}{c} 8+C-1 \\ C \end{array} \right)$ となる.

で，マルチモーダルな映像情報を扱う展開も可能である．これらについては，今後の検討課題とする．

7. むすび

視線の統計的解析に基づく映像視聴者の集中状態推定を目的として，線形システム集合によってシーンの顕著性変動を表現する scene-based saliency dynamics model (SSDM) を提案した．また，TV コマーシャルを用いた評価実験により，提案モデルの有効性を確認した．3.1 節で述べたように，本研究と [16] はそれぞれ顕著性変動，視線運動を中心に扱うという点で相補的なものであり，SSDM により記述されたシーンに対して, 視線ダイナミクスを考慮した視線特徴をデータドリブンに特徴選択して用いるという形で，両者の統合を行うことも可能である．前節の課題とあわせて，今後は映像・視線双方のダイナミクスに着目した，映像視聴状況のより高度なモデル化に取り組む．

謝辞 本研究の一部は，科学研究費補助金 特別研究員奨励費 24·5573 の補助を受けて行った．

文　献

[1] A. Yarbus: “Eye movements and vision”, Plenum (1967).

[2] J. Simola, J. Salojärvi and I. Kojo: “Using hidden Markov model to uncover processing states from eye movements in information search tasks”, Cognitive Systems Research, **9**, 4, pp. 237–251 (2008).

[3] S. Eivazi and R. Bednarik: “Predicting problem-solving behavior and performance levels from visual attention data”, Proc. Workshop on Eye Gaze in Intelligent Human Machine Interaction at IUI, pp. 9–16 (2011).

[4] 米谷, 川嶋, 加藤, 松山：“映像の顕著性変動モデルを用いた視聴者の集中状態推定”, 画像の認識・理解シンポジウム（MIRU）予稿集 (2012).

[5] P. Viola and M. Jones: “Robust real-time face detection”, International Journal of Computer Vision, **57**, 2, pp. 137–154 (2004).

[6] 山下, 藤吉：“特定物体認識に有効な特徴量”, 情報処理学会 研究報告 CVIM 165, pp. 221–236 (2008).

[7] C. Bregler: “Learning and recognizing human dynamics in video sequences”, Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 568–574 (1997).

[8] B. North, A. Blake, M. Isard and J. Rittscher: “Learning and classification of complex dynamics”, IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), **22**, 9, pp. 1016–1034 (2000).

[9] Y. Li, T. Wang and H.-Y. Shum: “Motion texture: a two-level statistical model for character motion synthesis”, ACM Transactions on Graphics (ToG), **21**, 3, pp. 465–472 (2002).

[10] G. Doretto, A. Chiuso, Y.-N. Wu and S. Soatto: “Dynamic textures”, International Journal of Computer Vision, **51**, 2, pp. 91–109 (2003).

[11] 川嶋, 三井, 松山：“動画像における時空間ダイナミクスのモデル化”, 画像の認識・理解シンポジウム（MIRU）予稿集, pp. 339–346 (2008).

[12] K. Ishiguro, T. Yamada and N. Ueda: “Simultaneous clustering and tracking unknown number of objects”, Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1–8 (2008).

[13] A. Chan and N. Vasconcelos: “Modeling, clustering, and segmenting video with mixtures of dynamic textures”, IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), **30**, 5, pp. 909–926 (2008).

[14] A. Ravichandran, R. Chaudhry and R. Vidal: “View-invariant dynamic texture recognition using a bag of dynamical systems”, Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1651–1657 (2009).

[15] 柳井：“一般物体認識の現状と今後”, 情報処理学会論文誌. コンピュータビジョンとイメージメディア, **48**, 16, pp. 1–24 (2007).

[16] R. Yonetani, H. Kawashima, T. Hirayama and T. Matsuyama: “Mental focus analysis using the spatio-temporal correlation between visual saliency and eye movements”, Journal of Information Processing, **20**, 1, pp. 267–276 (2012).

[17] R. Yonetani, H. Kawashima, and T. Matsuyama: “Multi-mode saliency dynamics model for analyzing gaze and attention”, Proc. Eye Tracking Research & Applications (ETRA), pp. 115–122 (2012).

[18] R. Goldstein, R. Woods and E. Peli: “Where people look when watching movies: do all viewers look at the same place?”, Computers in Biology and Medicine, **37**, 7, pp. 957–64 (2007).

[19] M. Dorr, T. Martinetz, K. Gegenfurtner and E. Barth: “Variability of eye movements when viewing dynamic natural scenes”, Journal of Vision, **10**, 10, pp. 1–17 (2010).

[20] P. Qvarfordt and S. Zhai: “Conversing with the user based on eye-gaze patterns”, Proc. of the SIGCHI Conference on Human Factors in Computing Systems (CHI), pp. 221–230 (2005).

[21] B. Brandherm, H. Prendinger and M. Ishizuka: “Interest estimation based on dynamic bayesian networks for visual attentive presentation agents”, Proc. ACM International Conference on Multimodal Interaction (ICMI), pp. 346–349 (2007).

[22] Y. Nakano and R. Ishii: “Estimating user’s engagement from eye-gaze behaviors in human-agent conversations”, Proc. International Conference on Intelligent User Interfaces (IUI), pp. 139–148 (2010).

[23] T. Hirayama, J. B. Dodane, H. Kawashima and

T. Matsuyama: “Estimates of user interest using timing structures between proactive content-display updates and eye movements”, IEICE Transactions on Information and Systems, **E-93D**, 6, pp. 1470–1478 (2010).

[24] A. Doshi and M. Trivedi: “Attention estimation by simultaneous observation of viewer and view”, Proc. IEEE Workshop on CVPR for Human Communicative Behavior Analysis (CVPR4HB), pp. 21–27 (2010).

[25] L. Fletcher and A. Zelinsky: “Driver inattention detection based on eye gaze-road event correlation”, International Journal of Robotics Research, **28**, 6, pp. 774–801 (2009).

[26] T. Hirayama, K. Mase and K. Takeda: “Detection of driver distraction based on temporal relationship between eye-gaze and peripheral vehicle behaviors”, Proc. International IEEE Conference on Intelligent Transportation Systems (ITSC) (2012).

[27] M. Mori, C. Miyajima, P. Angkititrakul, T. Hirayama, Y. Li, N. Kitaoka and K. Takeda: “Measuring driver awareness based on correlation between gaze behavior and risks of surrounding vehicles”, Proc. International IEEE Conference on Intelligent Transportation Systems (ITSC) (2012).

[28] B. Law, S. Atkins, A. Kirkpatrick and A. Lomax: “Eye gaze patterns differentiate novice and experts in a virtual laparoscopic surgery training environment”, Proc. Eye Tracking Research & Applications (ETRA), pp. 41–48 (2004).

[29] S. Eivazi, R. Bednarik, M. Tukiainen, M. von und zu Fraunberg, V. Leinonen and J. Jääskeläinen: “Gaze behaviour of expert and novice microneurosurgeons differs during observations of tumor removal recordings”, Proc. Eye Tracking Research & Applications (ETRA), pp. 377–380 (2012).

[30] L. Itti, C. Koch and E. Niebur: “A model of saliency-based visual attention for rapid scene analysis”, IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), **20**, 11, pp. 1254–1259 (1998).

[31] J. Harel, C. Koch and P. Perona: “Graph-based visual saliency”, Proc. Conference on Neural Information Processing Systems (NIPS), Vol. 19, pp. 545–552 (2007).

[32] A. Amini, R. Curwen and J. Gore: “Snakes and splines for tracking non-rigid heart motion”, Proc. European Conference on Computer Vision (ECCV), pp. 251–261 (1996).

[33] D. Cremers: “Dynamical statistical shape priors for level set-based tracking”, IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), **28**, 8, pp. 1262–1273 (2006).

[34] T. Cootes, G. Edwards and C. Taylor: “Active appearance models”, IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), **23**, 6, pp. 681–685 (2001).

[35] 有木：“DCT 特徴のクラスタリングに基づくニュース映像のカット検出と記事切り出し”, 電子情報通信学会論文誌, **J80-D-II**, 9, pp. 2421–2427 (1997).

[36] H. Kawashima and T. Matsuyama: “Multiphase learning for an interval-based hybrid dynamical system”, IEICE Transactions on Fundamentals, **E88-A**, 11, pp. 3022–3035 (2005).

[37] D. Panjwani and G. Healey: “Markov random field models for unsupervised segmentation of textured color images”, IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), **17**, 10, pp. 939–954 (1995).

[38] M. Cerf, J. Harel, W. Einhäuser and C. Koch: “Predicting human gaze using low-level saliency combined with face detection”, Proc. Conference on Neural Information Processing Systems (NIPS), pp. 1–8 (2007).

[39] C. Kanan, M. Tong, L. Zhang and G. Cottrell: “SUN: top-down saliency using natural statistics”, Visual Cognition, **17**, 6-7, pp. 979–1003 (2009).

[40] E. Kidron, Y. Schechner and M. Elad: “Cross-modal localization via sparsity”, IEEE Transactions on Signal Processing, **55**, 4, pp. 1390 –1404 (2007).

[41] Y.-F. Ma, X.-S. Hua, L. Lu and H.-J. Zhang: “A generic framework of user attention model and its application in video summarization”, IEEE Transactions on Multimedia, **7**, 5, pp. 907–919 (2005).

付　　録

1. 被験者実験の詳細

本節では，文献 [4], [16], [17] および本論文において用いられている被験者実験について，その詳細を示す．

1.1 実 験 条 件

実験において集中状態の異なる 2 条件を再現するために，被験者に対して以下の教示を行った．

条件 1（高集中条件）　映像を視聴してください．視聴後，映像にどの程度興味を持ったかを 7 段階で評価するアンケートに答えてください．

条件 2（低集中条件）　映像を視聴しながら 1000 から 7 を引き算しつづけ，その計算結果（1000, 993, ...）を音声で発話してください．正答率が悪い場合は，再実験になりますので注意してください．

条件 2 では，サブタスク（暗算タスク）に注意のリソースが割かれることによって映像視聴タスクへの集中度合いが低減することが期待される．さらに本実験では，被験者ができるだけサブタスクに集中するように,「正答率が悪い場合は，再実験になりますので注意してください」と教示した．また両条件ともに，被験

者には可能な限り視線を映像に向けるよう教示した．

条件 1 の試行から得られたアンケート結果の分布は平均値 4.74，標準偏差 1.18 であった．条件 2 におけるサブタスクの達成度は試行ごとに異なるが，計算結果の発話は各試行において最低 4 回は行われていた．

1.2 実験デザイン

10 名の被験者（いずれも 20〜30 代の男女）はそれぞれディスプレイ(注5)前に着席し，映像を視聴した．着席位置はディスプレイと被験者間の距離がおよそ 1m になるよう調整し，頭部の固定は行わなかった．ディスプレイ下には視線計測装置(注6)を設置し，被験者の視線を 30Hz で計測した．映像は 12 種類の TV コマーシャル（いずれも映像時間は 15 秒）を用い，音声とともに呈示した．

実験にあたって，12 種類の映像をそれぞれ 6 種類の映像からなる 2 グループ（A，B）に分割した．また 10 名の被験者は，それぞれ 5 名からなる被験者グループ（a, b）に分割した．各被験者は以下の手続きに従い，同一の映像を合計 2 回ずつ視聴した．

被験者グループ a

- **試行 1**: 映像グループ A — 条件 1
- **試行 2**: 映像グループ B — 条件 2
- **試行 3**: 映像グループ B — 条件 1
- **試行 4**: 映像グループ A — 条件 2

被験者グループ b

- **試行 1**: 映像グループ B — 条件 1
- **試行 2**: 映像グループ A — 条件 2
- **試行 3**: 映像グループ A — 条件 1
- **試行 4**: 映像グループ B — 条件 2

各試行において，映像グループ内での映像の再生順は被験者ごとにランダム化されている．また，試行 1 の前および試行 2, 3 の間には小休止を設定し，その際視線計測装置のキャリブレーションを行った．

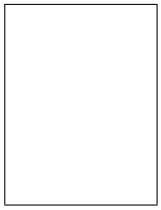
1.3 視線データに対する前処理

得られた視線データには，前処理として 0.5 sec の窓幅からなるメディアンフィルタを適用し，突発的ノイズの抑制および瞬きによる短時間データ欠損の補間を行った．前処理後，閉瞼や画面外注視によるデータ欠損は全体のうち 23.6%であり，手法の評価にはこれらの欠損を除いた残りのデータを用いた．

(注5)：MITSUBISHI Diamondcrysta RDT262WH, 25.5 inch, W550 mm/H344 mm.

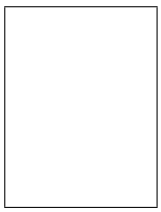
(注6)：Tobii X60 Eye Tracker. 頭部の移動可能な範囲は 400×220×300mm．本実験における計測誤差は平均 0.7° であった．

（平成 xx 年 xx 月 xx 日受付）



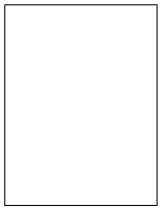
米谷　竜　（学生員）

2011 年京大大学院修士課程修了．現在，同大大学院情報学研究科博士後期課程在籍中．2012 年より日本学術振興会特別研究員（DC2）．視覚的注意モデル，視線解析の研究に従事．2010 年 ICPR IBM Best Student Paper Award．2012 年 MIRU 優秀学生論文賞．電子情報通信学会，情報処理学会学生会員．



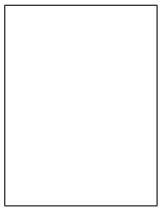
川嶋　宏彰　（正員）

2001 年京大大学院情報学修士課程了．2007 年より同大学院講師．2010 年から 2012 年までジョージア工科大学客員研究員（日本学術振興会海外特別研究員）．博士（情報学）．時系列パターン認識，ハイブリッドシステム，実世界インタラクションの研究に従事．2004 年 FIT 論文賞，2005 年船井ベストペーパー賞，2007 年 FIT ヤングリサーチャー賞．電子情報通信学会，情報処理学会，IEEE 会員．



加藤　丈和　（正員）

1997 年岡山大学工学部情報工学科卒業．2001 年同大大学院博士課程修了．博士（工学）．産業技術総合研究所特別研究員，和歌山大学システム工学科助手，同大学講師，情報通信研究機構特別研究員，京都大学情報学研究科特定研究員，同大学術情報メディアセンター特定准教授を経て，2012 年より同大学情報学研究科特定研究員．パターン認識，データマイニング，エネルギーの情報化の研究に従事．情報処理学会，電子情報通信学会，IEEE 各会員．



松山　隆司　（正員：フェロー）

1976 年京大大学院修士課程修了．京大助手，東北大助教授，岡山大教授を経て 1995 年より京大大学院電子通信工学専攻教授．現在同大学院情報学研究科知能情報学専攻教授．2002 年学術情報メディアセンター長，京都大学評議員，2005 年情報環境機構長．2008 年副理事．工博．画像理解，分散協調視覚，3 次元ビデオの研究に従事．最近は「人間と共生する情報システム」,「エネルギーの情報化」の実現に興味を持っている．1995 年 ICCV Marr Prize，2009 年文部科学大臣表彰科学技術賞（研究部門）ほか多数受賞．国際パターン認識連合，情報処理学会，電子情報通信学会フェロー．日本学術会議連携会員．

Abstract This article is the authors' version of Yonetani, Kawashima, Kato and Matsuyama: "Modeling Video Saliency Dynamics for Viewer State Estimation", IEICE Transaction, J96-D(8), pp.1675-1687, Aug. 2013 (http://search.ieice.org/bin/summary.php?id=j96-d_8_1675&category=D&year=2013&lang=J&abst). Videos used in this research including Figures 1, 4 and 7 were provided by courtesy of Panasonic Corporation. Eye movements that occur while humans are watching a video reflect their internal states as well as dynamic characteristics of the video scenes being watched. For a framework to analyze the relationship between the scene characteristics, eye movements and human states, we propose a novel model that describes the video scenes using primitive spatio-temporal patterns of video salient regions, which we refer to as the scene-based saliency dynamics model (SSDM). The SSDM introduces a set of linear dynamical systems to model the scenes (i.e., primitive patterns), and realizes statistical learning of eye movement features while considering the scene dynamics. Experimental results reveal the effectiveness of the SSDM by applying the model to estimate viewers' attentive states when watching general TV commercials.

Key words saliency dynamics, saliency map, eye movements, viewer state estimation