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†　　川嶋　宏彰†　　加藤　丈和††　　松山　隆司†

† 京都大学 大学院情報学研究科　〒 606-8501 京都市左京区吉田本町
†† 京都大学 学術情報メディアセンター　〒 606-8501 京都市左京区吉田本町
E-mail: †{yonetani,tkato}@vision.kuee.kyoto-u.ac.jp, ††{kawashima,tm}@i.kyoto-u.ac.jp

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”, 第 15 回画像の認識・理解シンポジウム (MIRU2012), Aug. 2012 の著者バージョンです．図 1, 4 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．This article is the authors' version of Yonetani, Kawashima, Kato and Matsuyama: ”Modeling Video Saliency Dynamics for Viewer State Estimation”, in MIRU, Aug. 2012. Videos used in this research including Figures 1 and 4 were provided by courtesy of Panasonic Corporation.
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，映像のシーン特性を考慮した視線解析に基づく視聴者の集中状態推定を目的とし，映像中の顕著領域が織りなす変動（顕著性変動）を線形システムを用いて表現する 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）ごとに視線特徴の統計的学習を行うことが可能となる．本稿では，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モデル化された顕著性変動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. はじめに

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，同一画像を対象とした場合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，視線情報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，視線計測技術の発達にともない，視線特徴の統計的学習を用いた人間の状態推定が各種提案されている [2]〜[4]．本研究では，人間がニュースや TV コマーシャルといった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．

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 [5]．したがって，映像視聴中の視線情報を解析する上では，映像中のシーンの特性を理解することが重要となる．画像・映像の認識および理解は，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，認識において実用レベルの技術が提案されている [6], [7] ほか，近年は一般物体認識にスポットが当てられている [8]．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 [9]〜[16]．

しかしながら，本研究で扱うような一般映像において認識されたシーンの特性を視線情報と紐づけて考える場合，シーンの持つ多様性が問題となる．一般物体認識においてしばしば問題となるように，一般映像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，カメラワークによりその見えは大きく変化する．したがって，一般映像に対する視線情報の統計的解析には，このような多様性を持つシーンに対してそれぞれ視線特徴を学習しなければならないという困難さが含まれる．

そこで本研究では，映像の意味的情報を捨象した特性として，顕著性（saliency）に着目する．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．このような複数の顕著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の種類に分類して用いることで，映像の多様性を許容しつつシーン（変動パターン）ごとに視線特徴の統計的学習を行うことが可能になる．

本稿では顕著性変動のモデルとして，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（図 1）．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．本研究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

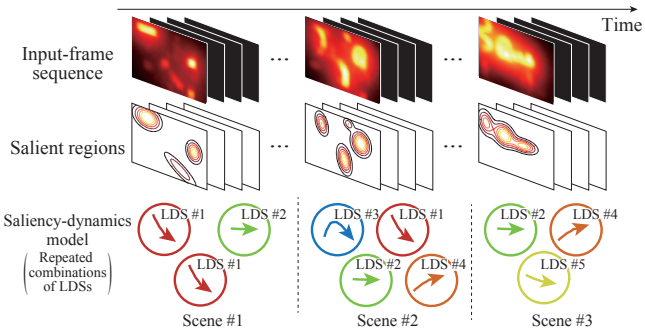

図 1　Scene-based saliency dynamics model.

シーンを記述する．すなわち SSDM 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集合により記述することで，映像全体の顕著性変動を表現する．

SSDM のモデル推定にあたっては，与えられた映像から顕著領域の変動パターンを抽出し，線形システムを同定することで，映像を構成するシーンを獲得する．この際，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映像のシーン分割を行うことで，映像全体として誤差の小さいモデルを推定する．

つづく 2. 節では，SSDM による顕著性変動のモデル化について，関連研究も交えてより具体的に説明する．3. 節では SSDM のモデル推定法を提案する．4. 節では，モデル化された顕著性変動と視線運動を用いた映像視聴者の集中状態推定法を提案する．5. 節では，TV コマーシャル映像に対する提案モデルの適用例，および提案モデルを用いた集中状態推定の精度評価を示す．

2.　SSDM を用いた顕著性変動のモデル化

2.1　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 [17] などで提案されている顕著性マップ（saliency map）を用いる．[17] 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が得られる．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 $i_t : \mathbb{N}^2 \to \mathbb{R}^+$ と表記する(注1)．

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．

人間の状態推定に対する本研究のアプローチは，映像のシーン記述において顕著性変動に着目し，シーン（変動パターン）ごとの視線特徴を統計的に学習することで状態推定を行うというものである．このような顕著性変動に基づく視線解析では，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)　与えられたシーンに対して視線情報をどのように解析するか（視線特徴をどのように設計するか）

このうち (B) の問題を中心に扱った研究として [4] がある．[4] では映像のシーンをフレーム中の顕著領域数（単数，複数）および領域の動き（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て用いるというアプローチで集中状態の推定を行っている．これに対して本研究は特に (A) の問題に着目し，映像のシーンを，顕著領域数およびそれら領域のダイナミクスに基づいてより詳細に記述する．このとき領域のダイナミクスを線形システムによってモデル化することで，領域の位置変化のみならず形状や顕著度の変化も同時に表現することが可能である．

2.2　Scene-based saliency dynamics model

顕著性変動のモデル化では，フレーム中に現れる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，シーン系列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って表現することで，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のモデルとして定義する．すなわち本研究で提案する SSDM は，線形システム集合の切り替わりによって，顕著領域数および領域ダイナミクスが変化する顕著性変動を表現するモデルである．

SSDM の説明にあたって，まず映像がシーン時区間系列 $\mathcal{I} = (I_1, \ldots, I_K)$ 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 3.3 節で提案する）．時区間 $I_k = [b_k, e_k]$ における顕著性マップ $\{i_t \mid t \in I_k\}$ はそれぞれ C_k 個の顕著領域を含むものとし，フレーム t における c 番目（$c \in \{1, \ldots, C_k\}$）の顕著領域の特徴が列ベクトル $\boldsymbol{\theta}_t^{(c)} \in \mathbb{R}^N$ で表現され，この領域の変動パターンは $\Theta^{(c,k)} = (\boldsymbol{\theta}_{b_k}^{(c)}, \ldots, \boldsymbol{\theta}_{e_k}^{(c)}) \in \mathbb{R}^{N \times (e_k - b_k + 1)}$ とベクトル系列で表されるものとする．すると，時区間 I_k における顕著性変動は，顕著領域の変動パターンの集合 $\boldsymbol{\Theta}_k = \{\Theta^{(1,k)}, \ldots, \Theta^{(C_k,k)}\}$ によって表現できる．

SSDM では，このような顕著性変動パターンの集合を用いてシーン $\boldsymbol{w}_k$ 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，線形システム集合 $\mathcal{D} = \{D_1, \ldots, D_S\}$ を考える．そして，シーン $\boldsymbol{w}_k$ によって $\mathcal{D}$ の重複有りサブセットを表現する．すなわち，$\boldsymbol{\Theta}_k$ の各要素に対して $\mathcal{D}$ のうちいずれかの線形システムを同

(注1)：顕著性マップは [18] 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．

定することにより，$\boldsymbol{w}_k$ を S 次元ベクトル

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

として定義する．ここで，w_{sk} は D_s の同定された変動パターンの数である（したがって $\sum_s w_{sk} = C_k$）．シーンの導入により，映像全体の顕著性変動が $\mathcal{W} = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_K)$ としてモデル化できることになる．

コードブックの各要素となるシステム $D_s \in \mathcal{D}$ 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．

$$\boldsymbol{z}_t = F^{(s)}\boldsymbol{z}_{t-1} + \boldsymbol{g}^{(s)} + \boldsymbol{v}_t^{(s)}. \tag{2}$$

$\boldsymbol{z}_t$ は時刻 t における領域の状態（すなわち $\boldsymbol{z}_t = \boldsymbol{\theta}_t^{(c)}$），$F^{(s)}$ は遷移行列，$\boldsymbol{g}^{(s)}$ はバイアス，$\boldsymbol{v}_t^{(s)}$ はガウス分布 $\mathcal{N}(0, Q^{(s)})$ によってモデル化されるノイズである．このように D_s は，$F^{(s)}, \boldsymbol{g}^{(s)}, Q^{(s)}$ をパラメタに持つ．

2.3 映像ダイナミクスのモデル化

ここで，いくつかの関連研究を紹介する．提案モデルのように対象を線形システム集合によって表現するモデルとして，bag of dynamical systems [16] がある．これは統計的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，時空間ボクセルとして表現される映像全体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．本研究で提案する 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 [9]〜[11]．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 [14] がある．[14] 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 3. 節）．[14] 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．

(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．

2.4 顕著領域のパラメタ表現

提案モデルを導入するにあたって，各顕著領域が持つ特徴（位置や形状，顕著度分布）パラメタ $\boldsymbol{\theta}_t^{(c)}$ 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 [19] やレベルセット法 [20] が，またテクスチャダイナミクスの表現手法としては dynamic textures [12], [16] 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 [13] や active appearance model (AAM) [21] 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，モデル化精度が手動での初期値学習に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 i_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 $\boldsymbol{\mu}_t^{(c)}, \Sigma_t^{(c)}, \phi_t^{(c)}$ 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 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 $(\boldsymbol{\mu}_{b_k}^{(c)}, \ldots, \boldsymbol{\mu}_{e_k}^{(c)}), (\boldsymbol{\sigma}_{b_k}^{(c)}, \ldots, \boldsymbol{\sigma}_{e_k}^{(c)}), (\phi_{b_k}^{(c)}, \ldots, \phi_{e_k}^{(c)})$ を考える（$\boldsymbol{\sigma}_t^{(c)} \in \mathbb{R}^3$ は $\Sigma_t^{(c)}$ 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 $(\bar{\mu}_{b_k}^{(c)}, \ldots, \bar{\mu}_{e_k}^{(c)}), (\bar{\sigma}_{b_k}^{(c)}, \ldots, \bar{\sigma}_{e_k}^{(c)}), (\bar{\phi}_{b_k}^{(c)}, \ldots, \bar{\phi}_{e_k}^{(c)})$ を領域の特徴として用いる（$\bar{\phi}_t^{(c)}$ は標準化のみ行う）．すなわち，変動パターン $\Theta^{(c,k)}$ 中の要素 $\boldsymbol{\theta}_t^{(c)}$ は，これらを用いて $\boldsymbol{\theta}_t^{(c)} = (\bar{\mu}_t^{(c)}, \bar{\sigma}_t^{(c)}, \bar{\phi}_t^{(c)})^{\mathrm{T}}$ として定義されるものとする．

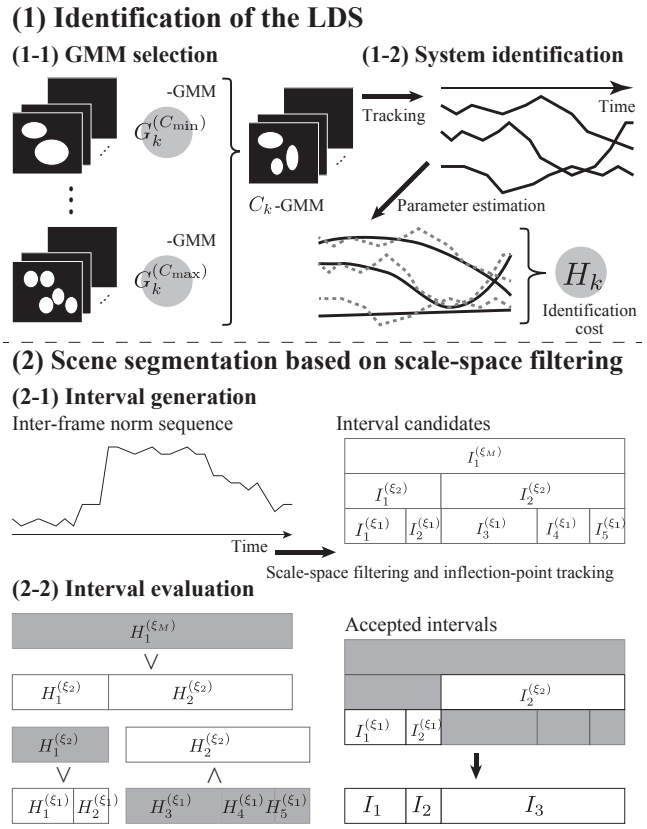

図 2　SSDM のモデル推定の流れ.

3. SSDM のモデル推定

3.1 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モデル化する．SSDM のモデル推定は，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 $\mathcal{D}$ の生成により行われる．

2.2 節で述べた通り，シーン $\boldsymbol{w}_k$ は $\mathcal{D}$ の重複ありサブセットとして，顕著領域数および領域のダイナミクスに基づいて記述される．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）線形システムが同定されるように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り，結果としてこれらは鶏と卵の問題になる．2.3 節で紹介した [14] の枠組みは，先に主体数を決めたのちにダイナミクスのパラメタ同定を行うため，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分割（主体数の決定）を行うことは難しい．

これに対して本研究では，多重解像度解析の枠組みを導入した，階層的なシーン分割推定法を提案する．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 コンポーネント数の推定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 $\mathcal{D} = \{D_1, \ldots, D_S\}$ 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 $\boldsymbol{w}_k = (w_{1k}, \ldots, w_{Sk})^{\mathrm{T}}$ 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 3.2 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 3.3 節では，線形システム同定コストを用いて多重解像度解析の枠組みでシーン分割を行う手法を述べる．3.4 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．

3.2 線形システム集合の同定

シーン時区間 $I_k = [b_k, e_k]$ における領域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の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 $(i_{b_k}, \ldots, i_{e_k})$ に対して，あらかじめ取りうるコンポーネント数の範囲 $\{C_{\min}, C_{\max}\}$ を定める．そして，2.4 節で述べた方法により各フレームの顕著性マップ i_t についてコンポーネント数 $C_{\min}, \ldots, C_{\max}$ の GMM を当てはめる．この際，コンポーネント数 C の GMM によって i_t をモデル化する際のコスト（GMM 化コスト）$g_t^{(C)}$ を，赤池情報量基準（Akaike information criterion; AIC）によって定める．すると，顕著性マップ系列 $(i_{b_k}, \ldots, i_{e_k})$ をコンポーネント数 C の GMM の系列としてモデル化する際の GMM 化コスト $G_k^{(C)}$ は，各フレームにおける GMM 化コストの和 $G_k^{(C)} = \sum_{t=b_k}^{e_k} g_t^{(C)}$ として表現できる．

この GMM 化コスト $G_k^{(C)}$ を最小にする C を，時区間 I_k におけるコンポーネント数 C_k として定める．

線形システムの同定（図 2 (1-2)）

引き続き変動パターン $\Theta^{(c,k)}$ に対して線形システムの同定を行う．ここでは，$\Theta^{(c,k)}$ に対して同定されるシステムを $D^{(c,k)}$ と表記する．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パラメタ $F^{(c,k)}, \boldsymbol{g}^{(c,k)}$ を推定すればよい．ただし，$\Theta^{(c,k)}$ 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，$A = [F^{(c,k)} \mid \boldsymbol{g}^{(c,k)}]$ に関する正則化項 $||A||_{\mathrm{F}}^2$ 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，リッジ回帰における正則化パラメタは $||A||_{\mathrm{F}}$ が一定閾値以下になるように設定する．パラメタ A が推定されると，初期値 $\boldsymbol{\theta}_{b_k}^{(c)}$ を与えることで変動パターンの生成が可能になる．これを用い，$\Theta^{(c,k)}$ から推定パラメタを用いて変動パターンを生成し，$\Theta^{(c,k)}$ 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 $Q^{(c,k)}$ を得る．そして，$\mathcal{N}(0, Q^{(c,k)})$ 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 $h^{(c,k)}$ が計算できる．

こうして，時区間 I_k における C_k 個の変動パターンについて，線形システム集合 $\{D^{(k,1)}, \ldots, D^{(k,C_k)}\}$ 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 $H_k = \max_c h^{(c,k)}$）を与える．

3.3 多重解像度解析に基づく映像のシーン分割

多重解像度解析によるシーン候補の生成（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あるいは類似フレームのクラスタリングによるアプローチがある（たとえば [22]）．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形成しておき，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 i_{t-1}, i_t 間の差分 $f_t = |i_t - i_{t-1}| \in \mathbb{R}^+$ 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 $\boldsymbol{f}$ に対してスケール $\{\xi_1, \ldots, \xi_M\}$ $(\xi_{m-1} < \xi_m)$ 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）．

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

このとき，波形集合 $\{\boldsymbol{f}^{(\xi_1)}, \ldots, \boldsymbol{f}^{(\xi_m)}\}$ において平滑化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，ここではスケール $\{\xi_1, \ldots, \xi_M\}$ を，すべてのスケール変化 $\xi_m \to \xi_{m-1}$ 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_M は変曲点を持たないスケールとする．

以下では，各スケールで得られた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_m において形成された時区間候補を $\hat{\mathcal{I}}^{(\xi_m)} = (\hat{I}_1^{(\xi_m)}, \ldots, \hat{I}_{K_{\xi_m}}^{(\xi_m)})$ と表記する．3.2 節で述べた手法によって，$\hat{I}_k^{(\xi_m)}$ 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 $H_k^{(\xi_m)}$ 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_m における時区間候補 $\hat{I}_k^{(\xi_m)}$ と同一区間に定義されるような，スケール ξ_{m-1} における $\hat{\mathcal{I}}^{(\xi_{m-1})}$ の部分系列 $\hat{\mathcal{I}}^{(\xi_{m-1})}|_{(j,j+l)} = (I_j^{(\xi_{m-1})}, \ldots, I_{j+l}^{(\xi_{m-1})})$ を考える．このとき，当該区間を時区間（系列）$\hat{I}_k^{(\xi_m)}$，$\hat{\mathcal{I}}^{(\xi_{m-1})}|_{(j,j+l)}$ のいずれで表すか（すなわち当該時区間において分節化をおこなうかどうか）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 $H_k^{(\xi_m)} > \sum_{j'=j}^{j+l} H_j^{(\xi_{m-1})}$ ならば分節化を行う．このような評価を，最高スケール ξ_M における時区間 $\hat{\mathcal{I}}^{(\xi_M)}$ 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．

3.4 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 C_k 個の要素よりなる線形システム集合 $\{D^{(k,1)}, \ldots, D^{(k,C_k)}\}$ 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，S 個の線形システムからなるコードブック $\mathcal{D} = \{D_1 \ldots, D_S\}$ を生成することができる．そして，$\{D^{(k,1)}, \ldots, D^{(k,C_k)}\}$ において D_s に分類されたシステム数を w_{sk} 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 $\hat{S}$ 個の線形システム $\{D_1 \ldots, D_{\hat{S}}\}$ 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，$S < \hat{S}$ 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 [16]．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_s から生成された固定長 l の変動パターン $\Theta_s \in \mathbb{R}^{3\times l}$ 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．Θ_s を行ベクトルに分解したものを $\Theta_s = ((\boldsymbol{\eta}_1^{(s)})^{\mathrm{T}}, (\boldsymbol{\eta}_2^{(s)})^{\mathrm{T}}, (\boldsymbol{\eta}_3^{(s)})^{\mathrm{T}})^{\mathrm{T}}$ と表す（$\boldsymbol{\eta}_p^{(s)}$ は長さ l の行ベクトル）．このとき，パターン Θ_i, Θ_j の非類似度 $Z(\Theta_i, \Theta_j)$ を式 (4) により定義する．

$$Z(\Theta_i, \Theta_j) = 1 - \sum_{p=1}^{3} \omega_p \mathrm{ZNCC}(\boldsymbol{\eta}_p^{(i)}, \boldsymbol{\eta}_p^{(j)}) \qquad (4)$$

ただし，ZNCC は正規化相互相関，ω_p $(\sum_p \omega_p = 1)$ 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，コードブックのシステム数 S を決定する．具体的には，最近傍に存在する線形システム（集合）同士を一つのクラスタに併合するという処理を，クラスタ数が 1 になるまで繰り返す．なお線形システム集合間の非類似度は，各集合に属する変動パターンの全組み合わせから得られる非類似度の最大値により定義する．

階層クラスタリングの過程において，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し，AIC の和を算出することによって，クラスタリングコストを計算する．このコストによって，最終的なクラスタ数を決定する．ただし，このような線形システムのクラスタリングにおいては情報量基準の自由度の大きさを適切に評価することが難しい [23]．そこで本研究では文献 [23] に倣い，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 S として手動で定める．コードブック $\mathcal{D} = \{D_1 \ldots, D_S\}$ 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．

4. 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．

4.1 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，人間がそれを視聴している状況を想定する．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，映像に対する視聴者の集中の高さを推定する問題である [4]．

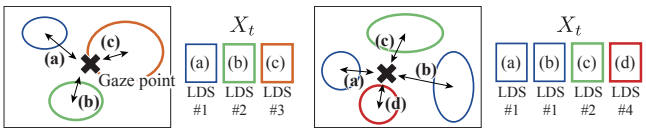

図 3　注視点と顕著領域間の空間的構造．

4.2 視線特徴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．[4] 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注視点移動の頻度）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けるシーンは，線形システム集合によって記述されるものであり，[4] で想定されたシーンと比較してより多様性を持つことから，このようなトップダウンの特徴設計を行うことは困難であると考えられる．

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，注視点と顕著領域間の空間的構造を用いる（図 3）．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 $(1, \ldots, S)$ が小さな順に並びかえられているものとする（ID が重複する場合 $\boldsymbol{\mu}_t^{(c)}$ はラスター順に並ぶものとする）．このとき，時刻 $t \in I_k$ における注視位置 $\boldsymbol{x}_t \in \mathbb{R}^2$ を用いて，視線特徴 X_t を以下のように定義する．

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

X_t は “どのようなダイナミクスを持つ領域が注視されるか” を表しており，シーン特性を反映していると言える．また，全シーンについて視線特徴が式 (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．

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

5. 実　　験

a）**一般映像に対する提案モデルの適用**

提案モデルを一般映像に適用した例を示す．本研究で

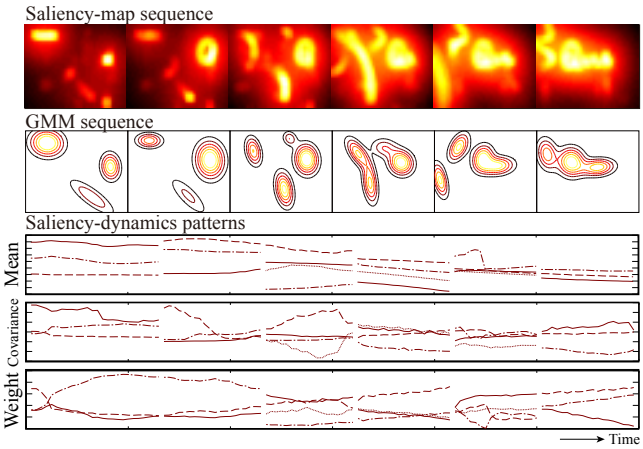

図 4　顕著性マップ系列より抽出された顕著性変動パターン．1 段目: 入力の顕著性マップ系列（各シーンの第 1 フレーム），2 段目: GMM のフィッティング結果，3〜5 段目: 顕著性変動パターンを表す．

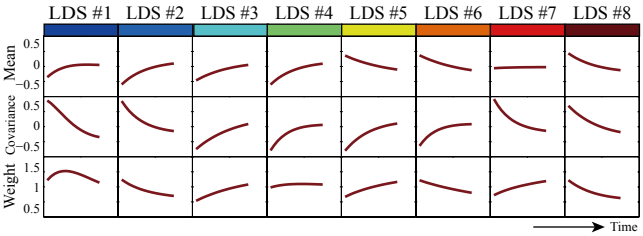

図 5　生成されたコードブック．各 LDS の色は図 6 と対応する．

は, [4] と同様の TV コマーシャル 12 種類を用いた．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．3.2 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 $C_{\min} = 1, C_{\max} = 8$ に設定した．また SSDM のモデル推定に関して，3.4 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_p は $\omega_1 = \omega_2 = \omega_3 = 1/3$ に設定した．

提案モデルの適用結果を示す．図 4 に入力の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 個の領域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，クラスタリングコストの変化に基づいて $S = 8$ のコードブックを生成した（図 5）．図 4 における変動パターンのシーン表現を図 6 に示す．本実験では合計 142 種類のシーンが獲得された．

b）**提案モデルを用いた集中状態推定**

提案手法の性能を，集中状態推定を通して確認する．映像コンテンツおよび視線データには文献 [4] と同様のデータセットを利用した．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．

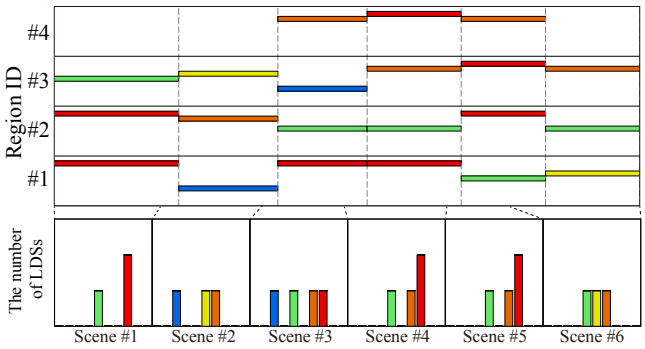

図 6　変動パターンのシーン表現．上段: 顕著領域に割り当てられた線形システム．色，長方形の位置はシステムの ID を表す．下段: 各シーンの顕著性変動．縦軸は各システムの出現回数，横軸はシステムの ID を表す．

表 1　実 験 結 果

Method	Duration	PERCLOS	[4]	**Proposed**
Accuracy (%)	53.8	65.0	78.2	**85.4**

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　映像を視聴しつつ，暗算タスクを遂行する．
これらはそれぞれ映像視聴に対する高集中，低集中状態に対応している．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（実験の詳細は [4] を参照）．

提案手法の精度評価は，データ（10 名 ×12 映像 ×2 条件 =240 試行，各集中状態について 120 試行）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，投票された視線特徴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．

c）**結果および討議**

表 1 に推定精度を示す．比較手法として, [4] の提案手法およびそのベースラインである顕著領域の注視時間（表中 Duration）および percentage of eyelid closure（同 PERCLOS）による推定を掲載した．実験の結果，従来研究と比較して精度の改善が確認された．

[4] はシーンを顕著領域数（単数，複数）およびその動き（静止，移動）のみに基づいて記述しており，顕著領域の個数の増加にともなう注視対象の分散の程度や，動き方の異なる顕著領域に対してどちらに注視が向きやすいかといった注視の傾向は考慮されない．これに対して提案モデルでは，顕著領域の個数も含めてシーンの記述を行うほか，顕著領域の動き方の違いをコードブックより割り当てられる線形システムの違いとして表現する．そして，あるシーンにおいてどのダイナミクスに対して注視が向きやすいかを，式 (5) に示した注視点と顕著領域間の空間的構造を用いて特徴化している．結果として, [4] と比較してより効果的なシーン表現および視線特徴抽出ができ，精度向上に繋がったと考えられる．

一方で両手法には，学習セット中に存在しない（視線

特徴の学習されていない）シーンに対する推定を行うことができないという共通の限界が存在する．そのため，提案モデルを利用する際にはシーンの種類数が問題となる．本実験で用いた映像からは合計 142 種類のシーンが獲得された．しかし，S 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 $\left(\begin{smallmatrix} S+C-1 \\ C \end{smallmatrix} \right)$ 種類になりうる．本研究では $S=8$，$3 \leqq C \leqq 5$ と推定されており，考えられうるシーンは $\sum_{C=3}^{5} \left(\begin{smallmatrix} 8+C-1 \\ C \end{smallmatrix} \right)$=1242 種類となる．すなわち，本実験において視線特徴の学習されたシーンは全体の 11.4%となる．提案手法はコードブック $\mathcal{D}$ を学習することで未知映像のシーン表現も可能であるが，未知映像に対する集中状態推定を行う場合，このような学習済みシーンの割合が重要となる．実際にはシーンの出現頻度には偏りが存在するためこの値は必ずしも有効な指標ではないが，視線特徴が学習されるシーンをより多くするためには，推定精度が保証される範囲で S を小さくすることが有効であると考えられる．たとえば $S=5$ の場合，推定精度は 81.1%と低下するが，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．その一方で，集中状態推定の精度向上や視線特徴学習の容易さを目的とする場合，推定精度や学習データに含まれるシーンの割合を最大化するようコードブックを定める方法も考えられる．これについては今後の検討課題とする．

なお，本研究では映像の持つ視覚的顕著性に着目し，顕著領域の抽出では典型的な顕著性マップ [17] を用いた．ただし，顕著領域は必ずしも視覚的顕著性のみに基づいて抽出する必要は無く, [24] や [25] のように顕著性に特定物体検出を組み合わせて用いることも可能である．

6. 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．集中状態の推定にあたって，本研究で用いた視線特徴はフレームごとに得られるものである．一方，視線運動はそれ自体がダイナミクスを持つものであり, [4] で用いた視線特徴のように，ある時区間における視線運動の特性に着目することも重要である．2.1 節で述べたように，本研究と [4] はそれぞれ顕著性，視線を中心に扱うという点で相補的なものであり，提案モデルを用いて記述されたシーンに対してダイナミクスを考慮した視線特徴を統計的に学習するという形で，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”, IUI, pp. 9–16 (2011).
[4] R. Yonetani, H. Kawashima, T. Hirayama and T. Matsuyama: “Mental focus analysis using the spatio-temporal correlation between visual saliency and eye movements”, JIP, **20**, 1, pp. 267–276 (2012).
[5] 大野, 中谷, 中根：“マニュアルデザインにおける視線パターンと印象の関係”, HCG シンポジウム予稿集, pp. 37–44 (2011).
[6] P. Viola and M. Jones: “Robust real-time face detection”, IJCV, **57**, 2, pp. 137–154 (2004).
[7] 山下, 藤吉：“特定物体認識に有効な特徴量”, CVIM, pp. 221–236 (2008).
[8] 鷲見, 内田, 佐藤, 佐藤, 日浦, 福井, 馬場口：“パターン認識・メディア理解のグランドチャレンジ 5. パターン認識・メディア理解の 10 大チャレンジテーマ”, 信学会誌, **92**, 8, pp. 665–675 (2009).
[9] C. Bregler: “Learning and recognizing human dynamics in video sequences”, CVPR, pp. 568–574 (1997).
[10] B. North, A. Blake, M. Isard and J. Rittscher: “Learning and classification of complex dynamics”, TPAMI, **22**, 9, pp. 1016–1034 (2000).
[11] Y. Li, T. Wang and H. Shum: “Motion texture: a two-level statistical model for character motion synthesis”, ToG, **21**, 3, pp. 465–472 (2002).
[12] G. Doretto, A. Chiuso, Y. Wu and S. Soatto: “Dynamic textures”, IJCV, **51**, 2, pp. 91–109 (2003).
[13] 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”, MIRU 予稿集, pp. 339–346 (2008).
[14] K. Ishiguro, T. Yamada and N. Ueda: “Simultaneous clustering and tracking unknown number of objects”, CVPR, pp. 1–8 (2008).
[15] A. Chan and N. Vasconcelos: “Modeling, clustering, and segmenting video with mixtures of dynamic textures”, TPAMI, **30**, 5, pp. 909–926 (2008).
[16] A. Ravichandran, R. Chaudhry and R. Vidal: “View-invariant dynamic texture recognition using a bag of dynamical systems”, CVPR, pp. 1651–1657 (2009).
[17] L. Itti, C. Koch and E. Niebur: “A model of saliency-based visual attention for rapid scene analysis”, TPAMI, **20**, 11, pp. 1254–1259 (1998).
[18] J. Harel, C. Koch and P. Perona: “Graph-based visual saliency”, NIPS, **19**, pp. 545–552 (2007).
[19] A. Amini, R. Curwen and J. Gore: “Snakes and splines for tracking non-rigid heart motion”, ECCV, pp. 251–261 (1996).
[20] D. Cremers: “Dynamical statistical shape priors for level set-based tracking”, TPAMI, **28**, 8, pp. 1262–1273 (2006).
[21] T. Cootes, G. Edwards and C. Taylor: “Active appearance models”, TPAMI, **23**, 6, pp. 681–685 (2001).
[22] 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”, 信学論, **J80-D-II**, 9, pp. 2421–2427 (1997).
[23] H. Kawashima and T. Matsuyama: “Multiphase learning for an interval-based hybrid dynamical system”, IEICE Trans. on Fundamentals, **E88-A**, 11,

pp. 3022–3035 (2005).
[24] M. Cerf, J. Harel, W. Einhäuser and C. Koch: "Predicting human gaze using low-level saliency combined with face detection", NIPS, pp. 1–8 (2007).
[25] 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).