論　文

選択的注視に基づく複数対象の動作認識

和田　俊和[†]　　佐藤　正行[††]　　松山　隆司[†]

Multi-Object Behavior Recognition by Selective Attention

Toshikazu WADA[†], Masayuki SATO[††], and Takashi MATSUYAMA[†]

あらまし　本論文では，対象の切り出しが行われていない動画像(映像)から複数対象の動作識別と個数の認識を行う動作認識法を提案する．従来の動作認識法では，画像からの特徴抽出と，抽出された特徴の時系列解析，を順に行うボトムアップ処理が主流であった．このため，特徴抽出段階で画像から対象の切り出しを行うと，その処理が不安定であるために，動作認識システム全体が不安定になるという問題があった．本論文では複数動作を安定に識別するために，まず「選択的注視機構」を提案する．この機構は，系列の解析を行う非決定性有限オートマトン(NFA)の各状態に特定の画像領域(注目領域)を対応付けておき，この内部で特徴抽出(イベント検出)を行うものである．この方式では，注目領域外部の画像の変化の影響を受けないイベント検出，非決定性状態遷移による可能な全てのイベント系列の解析が行えるため，対象の切り出しが行われていない映像から複数の動作を同時に識別することができる．このNFAの活性な状態集合に，動作対象を表す色付きトークンを割り当て，状態遷移と共に伝搬させる「対象弁別機構」を導入することにより，動作の識別と同時に対象の個数の認識を行う動作認識システムを構成することができる．さらに，本システムを多視点映像に拡張する方法を示し，実験によって正確かつ安定に複数動作の認識が行えることを示す．

キーワード　複数対象の動作認識，トップダウン，ボトムアップ，注目領域，イベント検出，非決定性有限オートマトン

1. はじめに

動画像(映像)からシーン中での移動対象の動きを認識する問題は，何を認識するかに応じて，次の３つのレベルに分類することができる．

- 運動解析(対象の物理的な運動の解析)

対象の形，位置，姿勢およびそれらの時間的変化を３次元空間もしくは２次元画像平面上で解析する問題．

- 動作識別(対象の特性と物理環境に拘束された運動パターンの識別)

対象の特性およびそれを取り巻く物理的環境に拘束された運動(動作)を対象としたものであり，対象の運動パターンを動作クラスに識別する問題．

- 行為理解(対象の動作からの意図理解)

手話やジェスチャなど，意図に基づく対象の動作(行為)から，行為の元になった意図を求める問題．

本論文では，自動視覚監視システムにおける基本的機能として，動作識別を取り上げ，特に通常の監視シーンにおいて一般的に観測される複数対象が写された映像を用いる場合について検討を行う．動作識別を複数対象の動作に適用することは，単に複数動作を同時に識別するだけでなく，対象の個数を認識することも意味する．この意味で，本論文で扱う問題は，複数対象の「動作識別」と「個数の認識」の２つの部分問題から成る「複数対象の動作認識問題」と呼ぶことができる．

一般に，画像を用いた動作の識別・認識システムは，1) 画像からの特徴抽出，2) 抽出された特徴系列の解析，の２つの部分から構成される，従来の研究では，図1(a)に示すように，特徴抽出 → 特徴系列の解析，というボトムアップ解析を行うシステム構成が採用され，特徴系列の解析部では隠れマルコフモデル(HMM)が一般的に用いられてきた[1], [2], [4]．HMMは，セグメンテー ションされた時系列データを識別する手法と

[†] 京都大学大学院情報学研究科，京都府
Department of Intelligence Science and Technology, Graduate School of Informatics, Kyoto University, Yoshida-Honmachi, Sakyo, Kyoto, 606-01 Japan
[††] 京都大学大学院工学研究科，京都府
Department of Electronics and Communication, Graduate School of Engineering, Kyoto University, Yoshida-Honmachi, Sakyo, Kyoto, 606-01 Japan

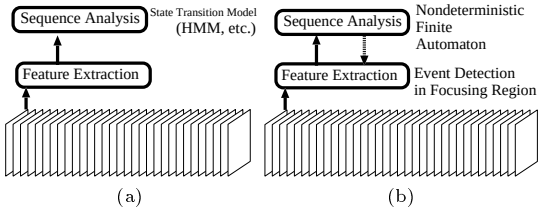

図1 動作識別システム (a): ボトムアップシステム, (b): ボトムアップ・トップダウンシステム.

Fig. 1 Behavior recognition system. (a): Bottom-up system, (b): Bottom-up and Top-down system.

しては強力な手法であるが，複数の時系列データが時空間で重畳している場合には，妥当な結果は得られない．したがって，従来法の枠組みで複数対象の動作を認識するには，特徴抽出段階で画像から対象の切り出しを行わなければならないが，この場合，切り出しの誤りが１回でも生じると，その誤りが特徴系列の解析部に蓄積され，システム全体の動作が不安定になる．このため，従来の研究では，画像特徴が比較的抽出しやすいlip reading [3] などの限られた用途でしか対象の切り出しは行われておらず，複数対象の動作認識問題は取り扱われてこなかった．

本研究では，画像からの対象の切り出しを行う事なく，映像から複数の対象の動作を同時に認識するシステムの構成法を提案する．これは，ドアや壁，階段など「剛体によって拘束された対象の動作を固定カメラで観測した際には，動作の各段階において画像上で変化が起きる場所が特定でき，既知の動作に対応する画像上の局所的変化の系列が全て検出できればその動作が起きたことが分かる」という考えに基づく手法である．例えば，ドアが閉じた部屋から人物が外に出ていく場面を撮影した画像では，人物の動作段階に応じてドアノブ，ドアの縁，…といった部分で順番に画像の変化が起き，そのような変化の系列が映像から検出できれば，'退室' という動作が起きたことが分かる．

この例のように，対象の動作段階と画像上の特定の領域(注目領域) 内での画像の変化 (イベント) を対応づけ，動作段階に対応するイベントを順次検出することによって，ある既知の動作が起きたことを判定する機構を「選択的注視機構」と呼ぶ．この機構は，動作の時系列を表す１次元状態系列から成る状態遷移モデルによって構成される．このモデルの各状態には，画像上の注目領域が対応づけられており，選択的注視機構は

- 注目領域内で画像からイベントの検出を行う．
- 検出されたイベントによって状態遷移を起す．
- 状態遷移後の活性化された状態に応じて注目領域を更新する．

という処理を反復する．この機構に画像系列を順次与えたとき，動作の開始を表す状態から終了を表す状態までの状態遷移が起きれば，当該動作が起きたと判定できる．この反復処理は，

ボトムアップ解析: イベント検出 → 状態遷移

トップダウン解析: 状態遷移 → 注目領域の更新

のように整理することができ，選択的注視機構は，従来のボトムアップ解析に加えて，トップダウン解析も行っていると言える (図1(b))．

選択的注視機構では，各時刻に活性化された状態を参照して注目領域を更新するため，状態遷移モデルとして，最適化計算後に過去の状態が確定するHMMは採用できない．また，イベント系列を安定に検出するために，検出されたイベントに対して，考えられ得る全ての状態への遷移が起きるようにしなければならない．このため，本研究では状態遷移モデルとして非決定性有限オートマトン (NFA) を用いる．

選択的注視機構は，1) 注目領域外部の画像の変化によって影響を受けないイベント検出，2) あるイベントに対して考えられ得る全ての状態への遷移を起すNFA，の2つによって，複数対象の動作を同時に解析することができる．また，複数の選択的注視機構を組み合わせることによって動作の識別を行うこともできる．しかしこの機構は，複数対象の動作識別機構であり，対象の個数を認識するためには，次の「対象弁別機構」を用いる必要がある．

選択的注視機構のNFAは，あるイベントに対して全ての可能な状態を同時に活性化するため，この機構だけでは動作対象の個数は求められない．この問題の解決策としては,「複数の活性化された状態集合のうち，時間的に連続した部分系列を１つの対象に対応するものと見なす」という方法が考えられる．しかし，このような部分系列の数は時刻とともに変化するため，対象の個数に関しても時系列的な解析を行う必要がある．「対象弁別機構」は，複数の活性化された状態集合のうち，時間的に連続した部分系列に同一の「色」を持つトークンを割り当て，状態遷移と共にそのトークンを伝搬させるという働きをする．この機構を用いることにより，動作の終了を表す状態に到達したトークンの色を数え上げ，対象の個数を数えることができる．

以上の選択的注視，対象弁別の両機構によって複数対象の動作認識システムが構成できる．

しかし，２次元画像は３次元世界の投影であるため，カメラの視線方向に沿った対象の移動に対応する画像上の変化は顕著ではなく，このような動作の場合には画像の変化から対象の動作段階を特定することは困難である．したがって，単一視点映像を用いただけでは，実用に耐え得る正確な動作認識システムは構築し難いと言える．この問題を解決するために，本論文では，上述の動作認識システムを拡張し，多視点映像を用いた動作認識システムの構成法も明らかにする．

実験では，室内を歩行する複数の人物の動作を，部屋への入室，退室の２クラスに分類するとともに，各動作を行っている人数を数える問題を取り扱い，本論文で提案したシステムにより，正確に複数対象の動作が認識できることを示す．

以下，２章では単一視点映像を用いた場合の動作認識法，３章で多視点映像への拡張法について述べ，４章では，実験結果について述べる．

2. 単一視点映像を用いた動作認識

ここでは，単一視点の映像を用いた場合の動作認識法について述べる．本論文で提案する動作認識システムは，ある動作が特定の動作クラスに属するか否かを判定する「動作同定モジュール」から構成されている．以下，動作同定モジュール，動作認識システム，注目領域の学習法について述べる．

2.1 動作同定モジュール

ここでは，選択的注視，対象弁別の両機構から成る動作同定モジュールの構成法について述べる．

2.1.1 選択的注視機構

この機構は，次の３つの要素から構成されている (図2)．

動作同定用 NFA: $(Q, q^0, \Sigma, \delta, F)$ の５つ組で表される．但し，

- Q:状態の有限集合,$\{q^0, q^1, \cdots, q^m, q^{rej}\}$. 状態の系列$(q^0, q^1, \cdots, q^m)$は対象の動作段階の時系列を表しており，$q^{rej}$ は入力が棄却されたことを表している．状態q^mへの遷移が起きたとき，入力は受理されるので，q^mはq^{acc}とも表す.
- q^0: 初期状態.
- Σ:イベントコード (後述) の集合.
- δ:状態遷移関数，$\delta(q, \sigma) : Q \times \Sigma \mapsto Q$ ．具体的には表1に示すような状態遷移表で与える．
- F:最終状態，$F = \{q^{acc}, q^{rej}\}$.

である．

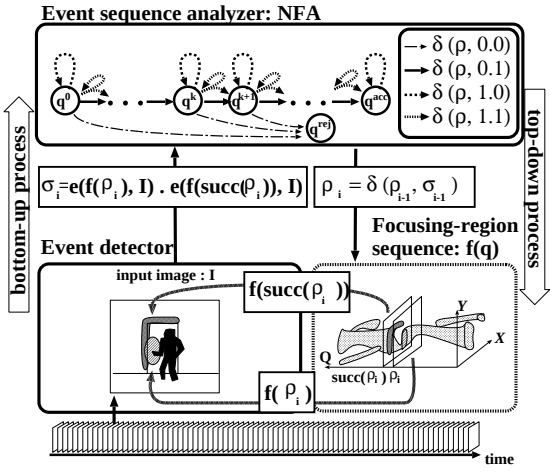

図2　選択的注視機構
Fig. 2　Selective Attention Mechanism.

現在の (活性) な状態を ρ_i で表すとき，$\rho_0 = q^0$ から出発し，イベントコード $\sigma_i (\in \Sigma)$ と現在の ρ_i に応じて活性な状態を $\rho_{i+1} = \delta(\rho_i, \sigma_i)$ によって更新する．

注目領域系列: あるクラスの動作に対して固有な画像上の変化が起きる場所 (注目領域) の系列．注目領域は，動作段階，すなわち NFA の各状態に対応付けられており，状態 q における注目領域は $f(q)$ で表す.

イベント検出器: 複数の注目領域内で入力画像の変化，すなわちイベントを検出し，複数のイベント検出結果を結合した「イベントコード」を出力する．

時刻tの入力画像$I(t)$に対して背景差分を適用して求められる，背景と画素値が大きく異なる画素の集合を$a(t)$，注目領域をf，イベント検出のための閾値を$\theta (0 < \theta < 1)$ とすると，イベント検出結果$e(f, I(t))$は次式で表される．

$$e(f, I(t)) = \begin{cases} 1, & f = \phi \quad or \quad \frac{|f \cap a(t)|}{|f|} > \theta \\ 0, & otherwise \end{cases} \quad (1)$$

である．但し，$|\cdot|$は画素数，ϕ は空の注目領域，検出結果1はイベントが検出されたこと，0は検出されなかったことを表す.

現在の活性な状態，および次に活性化され得る状態に対応する複数の注目領域内でのイベント検出結果を左から順番に‘.’でつなげたものがイベントコードであ

表1　$\rho_i = q^k$ における状態遷移表(イベントコード長が2の場合).

Table 1　State transition table at $\rho_i = q^k$ for event code length=2.

$e(f(q^k), I_i) \cdot e(f(q^{k+1}), I_i)$	ρ_{i+1}
$0 \cdot 0$	q^{rej}
$0 \cdot 1$	q^{k+1}
$1 \cdot 0$	q^k
$1 \cdot 1$	$\{q^k, q^{k+1}\}$

る．イベントコードの長さが2の場合，イベントコードの全体集合Σは$\{0 \cdot 0,\ 0 \cdot 1,\ 1 \cdot 0,\ 1 \cdot 1\}$になる．

対象同定は「現時の活性な状態を参照して注目領域を定め，それらを用いてイベント検出を行い，得られたイベントコードにより，NFAの状態遷移を起す」という手続きを繰り返し，q^{acc}への遷移が起きたかどうかを調べることによって実現される．イベントコード長が2の場合の具体的手続きを以下に示す．

Initialization: $i = 0,\ \rho_i = q^0$

Step1: $\sigma_i = e(f(\rho_i), I(t_i)) \cdot e(f(succ(\rho_i)), I(t_i))$(注1)

Step2: $\rho_{i+1} = \delta(\rho_i, \sigma_i)$

Step3: $i = i + 1$, goto **Step1**

この場合，表1に示すように，現在の活性な状態が$\rho_i = q^k$であるとき，イベントコードσ_iが$1 \cdot 1$になると，$\rho_{i+1} = \{q^k, q^{k+1}\}$になる．これが非決定性の状態遷移である．非決定状態遷移により，ある時刻に活性化されている状態は複数になり，これらの状態全てに，上述の**Step1**から**3**の手続きが適用される．さらに，初期状態の注目領域$f(q^0)$をϕにしておけば$f(q^0)$に対するイベントは常に検出されるため，q^0は常に活性化される．この結果，新たな動作が起るたびに，q^0を起点とする新たな状態遷移が起こる．

以上のように，選択的注視機構では，1)q^0を常に活性化しておき，2) あるイベントに関して考えられる全ての状態を活性化し，3) 遷移できなくなった(q^{rej}に遷移した)状態は不活性にする，という戦略を取ることにより，後戻りする事なく複数の動作を並列に同定することができる．

2.1.2　対象弁別機構

選択的注視機構では，1つの対象に対して複数の状態が同時に活性化されるため，このままでは対象の個数を認識することはできない．

対象の動作によって時間的に連続な画像の変化が起きると仮定すると，時間的に連続する注目領域は相互に類似していると言える．類似した注目領域内で同時にイベントが検出される可能性は高く，その結果1つの対象に対して，時間的に連続する状態集合が同時に活性化される．したがって，活性化された状態集合を，隣接する状態の部分系列に分解し，それらを動作対象と見なせば良いと思われる．

しかし，このような部分系列の数は時刻毎に異なり，正確な認識を行うには対象の個数に関しても時系列的解析を行う必要がある.「対象弁別機構」では，複数の活性化された状態集合のうち，1) 隣接した部分系列に同一の「色」を持つトークンを割り当てて，2) それらのトークンを状態遷移と共に伝搬させることによって，対象の個数に関する時系列的解析を行う．この機構を用いてq^{acc}に到達するトークンの色を数えることにより，動作対象の個数を認識することができる．

以下では，上述の色付きトークンの割り当てと，伝搬の方法について議論する．

[トークンの割り当て]

相互に隣接する「隣接状態集合」Cを，状態遷移関数δによって遷移可能な状態の集合とするとき，Cは次式を満足する．

$$\begin{aligned} \forall_{\sigma \in \Sigma}(\quad & \delta(\rho, \sigma) \neq q^{rej} \\ & \Rightarrow (\rho \in C \Rightarrow \delta(\rho, \sigma) \in C) \\ & \vee(\delta(\rho, \sigma) \in C \Rightarrow \rho \in C) \ ). \end{aligned} \tag{2}$$

活性な状態の全体集合Pは，上述の定義に基づいて，互いに素な隣接状態集合C^kに分解することができる．各C^kは異なる対象を表すものと考えているので，それぞれに異なる整数値z^kをトークンIDとして割り当てる．すなわち，活性な隣接状態集合とそれらに割り当てられるトークンは，$C^k \neq C^j \Rightarrow z^k \neq z^j$という「排他性条件」を満足するように割り当てる．

[トークンの伝搬]

1つの隣接状態集合に1種類のトークンIDしか割り当てないという制限を課すと，対象の個数が一定であっても，活性な隣接状態集合の個数，すなわち，トークンIDの個数は変化してしまう．したがって，複数の対象が近接して，複数の隣接状態集合が1つになった場合には，その隣接状態集合に複数のトークンID を割り当て，近接していた対象が離れて複数になった場合には，トークンIDの集合を各々に分配する，という操作を行えば，対象の個数に関する時系列的な解析が行える．以下，この具体的方法について述べる．

(注1)：$succ(\rho)$は，$\rho = q^i$のとき，q^{i+1}を返す後継者関数である．

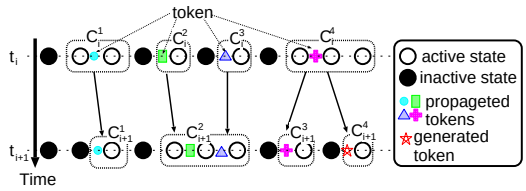

図3　リンクを介したトークンIDの伝搬.
Fig. 3　Token ID propagation via links.

時刻t_iとt_{i+1}の隣接状態集合を，それぞれC_i^j，C_{i+1}^kで表したとき，次式に示すように状態遷移関数δによって，C_i^jに含まれる状態から，C_{i+1}^kに含まれる状態への遷移が１つでも起きる場合，C_i^jからC_{i+1}^kへのトークンIDの伝搬が起きるものと考える．

$$\left(\bigcup_{\sigma \in \Sigma} \bigcup_{\rho \in C_i^j} \delta(\rho, \sigma) \right) \cap C_{i+1}^k \neq \phi, \tag{3}$$

この条件を満足するC_i^jとC_{i+1}^kの間には，トークンIDが伝搬する仮想的な路(リンク)が存在するものと見なす．このとき，トークンIDの伝搬と割り当ては，以下のように行う(図3)．

（１） 図3のC_i^1,C_i^2,C_i^3のように，C_i^jが単一のリンクを持つ場合，C_i^jに割り当てられたトークンIDの集合を単に伝搬させれば良い．図3のC_{i+1}^2のように，時刻t_{i+1}の隣接状態集合の一つに複数のリンクから異なるトークンIDが伝わってくる場合は，複数の対象が近接したため隣接状態集合が１つになった場合に相当する．

（２） 図3のC_i^4のように，C_i^jが複数のリンクを持つ場合には，近接していた複数の対象が離れたことに相当し，C_i^jに割り当てられたトークンIDの集合を適当に分割してt_{i+1}の隣接状態集合に伝搬させる．但し，排他性条件を満足させるため，異なるリンクを経由して同じトークンIDは流さない．また，トークンIDの集合を分割する際には，(4)のトークンIDの生成ができるだけ起きないように，均等に分割する．

（３） C_i^jがリンクを持たない場合，トークンの伝搬は起きない．これは誤って割り当てたトークンを削除することに相当する．

（４） トークンIDの伝搬終了後に，トークンが割り当てられていないC_{i+1}^kが存在した場合，新たなトークンIDを生成し，それらを割り当てる．これは，新たな対象が観測されたことを意味する．

但し，ここで述べた方法は，対象の個数を認識する方法であるので，C_i^iが複数のリンクを持つ場合，どのトークンをどのリンクに流すかを一意に決定できないことは問題ではない．

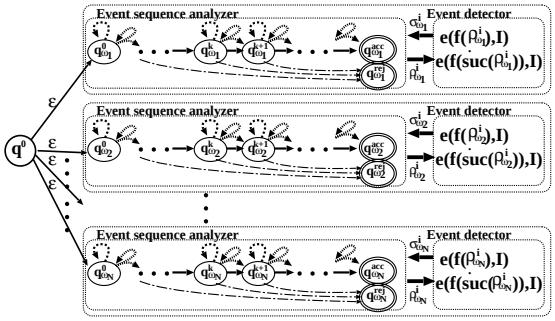

図4　動作認識システム
Fig. 4　Behavior classifier.

2.2　動作認識システム

複数の動作を識別する動作認識システムは，各動作クラスω_i $(i = 1, \cdots, N)$ に関する動作同定モジュールに，新たな初期状態q^0と，q^0から各同定機構の初期状態$q_{\omega_i}^0$へのϵ−遷移(注2)を付け加えることによって構成される(図4)．

2.3　注目領域の学習

動作同定モジュールのNFAは，対象の動作段階を表す状態の順序集合から構成され，個々の状態には，その状態が活性化されているときに画像上で変化が起きる部分，すなわち注目領域が対応づけられている．これら，状態と注目領域は，単一対象の動作を観測したデータが複数与えられているとき，以下のようにして求めることができる．

同じクラスに属する単一対象の動作を観測したn個の映像に対して，背景差分を適用し，変化領域系列のデータ$a^i(t)$ $(i = 1, \cdots, n)$が得られているものとする．これらのデータでは，対象の速度や大きさ，形などが異なるため，このままでは変化領域間の共通性を調べることができない．この問題を解決するため，以下のように，ある変化領域系列データ$a(t)$を基準として，他のデータの時間軸の正規化を行う．すなわち，次式で表される他の変化領域と$a(t)$との画像上での重なりの割合の総和を最大化する，単調増加関数τ^iを求める．

$$\int \frac{|a(t) \cap a^i(\tau^i(t))|}{|a(t) \cup a^i(\tau^i(t))|} dt, \tag{4}$$

(注2)：空入力に対する状態遷移

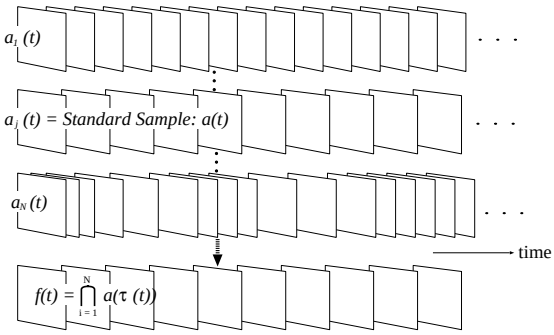

図5　共通変化領域の学習
Fig. 5　Learning a common anomalous region sequence.

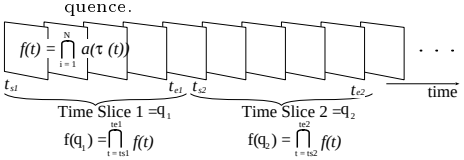

図6　注目領域の学習
Fig. 6　Learning a focusing region sequence.

但し，$|\cdot|$は画素数を表すものとする．この最適化問題は動的計画法を用いて解くことができる．

時間軸を正規化した変化領域系列から，各データで共通に変化が起きる画像領域$f(t)$の系列を以下のように求めることができる(図5)．

$$f(t) = \bigcap_{i=1}^{n} a^i(\tau^i(t)). \tag{5}$$

この系列を「共通変化領域系列」と呼ぶ．

共通変化領域系列は正規化された時間軸上で表現されている．この軸上の時間区間が，NFAの状態に対応し，各状態に対応する注目領域は共通変化領域系列から以下のように計算することができる(図6)．

NFAの状態qが時間区間$t_s \leqq t < t_e$に対応するとき，この状態の注目領域$f(q)$は次式で表される．

$$f(q) = \bigcap_{t=t_s}^{t_e-1} f(t) \tag{6}$$

時間区間の決定基準としては様々な方法が考えられるが，ここでは，できるだけ非決定性状態遷移が起きにくくなるように，1)注目領域が空ではなく，2)隣接する状態の注目領域ができるだけ異なるという2条件を満足するように時間区間を決定する．これは，まず$j = 0, q_j^i = t^i$とし，状態数が指定した個数になるまで$|f(q_j^i) \cap f(q_j^{i+1})|/|f(q_j^i) \cup f(q_j^{i+1})|$の値が最大になる$i$を求め，次式による状態の併合と注目領域の更新を行うという処理を繰り返せば良い．

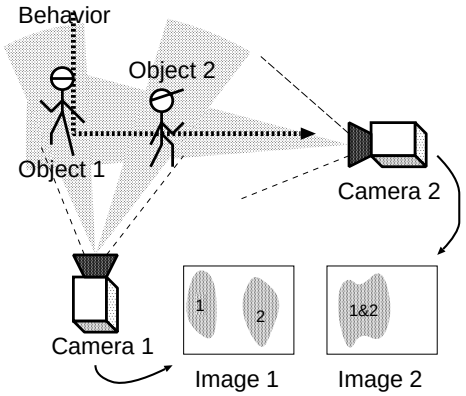

図7　複数映像を用いる効果．
Fig. 7　Effectiveness of multi-viewpoint images.

$$f(q_{j+1}^k) = \begin{cases} f(q_j^k), & k < i \\ f(q_j^k) \cap f(q_j^{k+1}), & k = i \\ f(q_j^{k+1}), & k > i \end{cases} \tag{7}$$

但し，$f(q_{j+1}^k) = \phi$となった場合は，$j = j - 1$として停止する．

3. 多視点映像への拡張

これまで述べてきた単一視点の映像を用いた動作認識システムでは，動作の種類によっては認識能力が低下する場合がある．これは，カメラの視線方向と対象動作の方向のなす角が小さい場合，動作が進行しても画像中における変化領域の位置・形状が変化しにくいため，動作段階の特定が困難になるためである．

この問題の解決法として，共通の視野を持ち視線方向の異なる複数(N_c)台のカメラを用いることによりお互いの弱い部分(視線方向への動作)を補う方法が考えられる．また，画像を複数枚用いることは，変化領域の3次元空間中での共通部分に注目領域を設定することと等価であり，対象の動作段階をより詳細に把握することができる(図7)．

複数カメラを使用するにあたって，次の2つ条件を仮定する．

- 各カメラは対象の動作が共通に撮影できるよう設置されている．
- カメラ間で画像の撮影時刻は同期している．

これらの条件は，いずれも複数の映像による動作認識を行うために必要であるが，カメラ間の空間的位置・

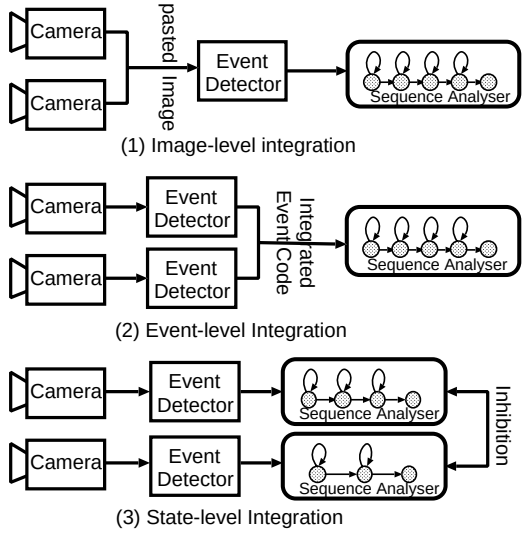

図8　3種類の情報統合法
Fig. 8　Three types of information integration.

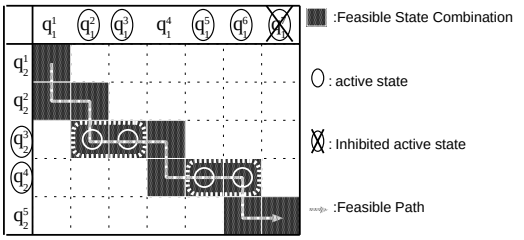

図9　状態の直積空間 ($N_c = 2$)
Fig. 9　State product space($N_c = 2$).

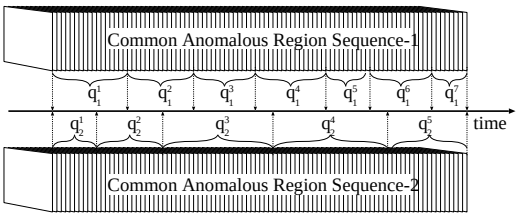

図10　許容経路の学習
Fig. 10　Learning a feasible path.

方位の正確なキャリブレーションは不要である．

複数の映像から1つの認識結果を得るためには，複数視点で得られた情報を何らかの方法で統合する必要がある．上述のように同期撮影された複数の映像を用いることができれば，前述の認識機構において，1) 映像，2) イベント，3) 状態，のいずれかのレベルで情報を統合することが可能である (図8)．　以下，これらの情報統合法について述べる．

3.1　映像レベルの統合法

まず最も単純な方法として，同一時刻に観測された複数の画像を張り合わせ，1枚の画像を作り，張り合わされた画像の系列 (映像) に対して前述の動作認識を行う方式が考えられる．これを「映像レベルの統合法」と呼ぶ．

この方式では，異なる視点から観測された画像を，すべて一様な入力として扱うため，図7に示したように，3次元空間の特定の領域における対象の変化をとらえるという効果は得られない．

3.2　イベントレベルの統合法

視点$c(c = 1, \cdots, N_c)$のカメラで観測した画像I_cから独立にイベント検出を行い，得られた複数のイベントコードe_cから1つのイベントコードe_{all}を生成し，それを1つのNFAの入力とする方式である．これを「イベントレベルの統合法」と呼ぶ．

e_{all}の具体的な計算法としては，変化領域の組に対応する3次元空間中の部分の変化をとらえるために，イベント検出結果の論理積$e_{all} = \cap e_c$を用いる方法が考えられる．この場合「すべてのイベント検出器でイベント検出が成功しなければならない」という強い条件を課しているため，対象の動作があらかじめ学習した動作から逸脱すると，統合されたイベント検出結果が未検出，すなわち0になり，イベント系列の検出が失敗しやすくなる．特に，カメラ台数N_cの増加に伴ってこの傾向は強くなる．

3.3　状態レベルの統合法

複数の動作識別システムの内部状態を統合することによって1つの認識結果を得る方式を「状態レベルの統合法」と呼ぶ．これは，動作識別システム内の動作同定モジュールを異なる視点間で同期して動作させ，あらかじめ学習しておいた「同時に活性化し得る共起性を有する状態組」に属さない状態を抑制する (不活性にする) ことにより，動作同定の曖昧さを減らす方式である．

視点$c(c = 1, \cdots, N_c)$の動作同定モジュールにおいて，対象の動作段階を表す状態集合$Q'_c = \{q_c^1, \cdots, q_c^{m_c}\}$には全順序関係が定義されている．したがって，図9のように，複数視点の動作同定モジュールの状態系列によって張られる直積空間$Q'_1 \times \cdots \times Q'_{N_c}$を各動作クラス毎に構成することができる．この直積空間内では，共起性を有する「許容状態組」は点になり，対象の動作を表す許容状態組の系列は一本の「許容経路」を形成する．状態遷移とトークンIDの伝搬をこの許容経路上で行うようにすれば，上述の共起性

を有しない状態の抑制が実現できる．この方法は,「各状態が対象の動作段階を表す」というこれまでの解釈を,「各許容状態組が対象の動作段階を表す」ととらえ直すことによって，選択的注視，対象弁別の各機構を多視点映像に拡張したことに相当している．

このように，状態間の共起性を利用することにより，単一視点映像による手法に比べて，３次元空間中でのより限定された部分の変化に注目することができるため，個々の対象の弁別性能が向上する．また，この方式では，許容経路を広げることによって，対象動作の空間的変動をある程度許容することが可能である．さらに，後述するように，この方式は，イベントレベルの統合法を特殊なケースとして含んでおり，より一般的な多視点映像への拡張法となっている．

3.3.1 許容経路の学習

ここでは，状態レベルの統合法における許容状態組，許容経路の学習法について述べる．

同じ動作データを基準として学習データの時間軸の正規化を行えば，異なる視点で撮影した映像の共通変化領域の時間軸を揃えることができる．また，動作同定モジュールの学習時には，個々の状態に対応する時間区間も求めることができる．したがって，同じ動作データを基準として時間軸を正規化しておけば，図10に示すように，各状態に対応する時間区間の重なりによって，異なる視点の動作同定モジュール間で状態に共起性があるか否かを判定することができる．この方法を用いることによって許容状態組，許容経路を学習することができる(図10の時間区間を用いた場合には，図9に示す許容経路が求められる)．

また，許容経路の拡張を行うには，一旦計算した許容経路を用いて，再度学習データに対する動作同定を行い，許容計路上のトークンIDの伝搬がとぎれる場合に，各動作同定モジュールの活性化された状態を参照して許容状態組を追加すればよい．

3.3.2 イベントレベルの統合法との関係

前述のイベントレベルの統合法では，１つの状態系列によって１つの動作クラスを表現しているため，各映像毎の注目領域系列は同一の時間区間に対応していると言える．これは，状態レベルの統合法においては，同じ時間区間に対応する状態の集合を持つ動作同定モジュールを複数用いる場合と等価である．このことから，イベントレベルの統合法が，状態レベルの統合法の特殊なケースとして表現できることが分かる．

4. 実　　験

ここでは，以上に述べた動作認識システムを実際に構成し，動作認識を行った結果を示す．

実験では，部屋に設置した２台のカメラによって撮影した２つの映像(白黒256 階調320×240，30[フレーム/秒]) から，画素値の変化量に関する閾値を20として背景差分を行い，この差分画像から人物の動作を「入室」「退室」に分類すると共にそれらの動作を行った人物の数を認識する問題を扱った．この実験では，ドアの開閉に要する時間を節約するため，ドアを解放した状態で実験を行った．また，入室後，退室前の対象の動作に制限を設けない場合も，出入り口を通過する前後で注目領域は学習できるので，動作認識も行える．しかし，今回の実験ではより長時間の映像を用いた認識を行わせるため，人物の動きを床面に描いたコース内に制限した．

学習用データとしては各動作クラスに対し，20組(40本) の映像を用い，これらから注目領域を学習した．入退室動作の原画像・変化領域 と注目領域の一例をそれぞれ図11，12に示す．これらの図から，動作に伴って変化する影等の人物そのものではない部分も注目領域の一部に含められていることが分かる．

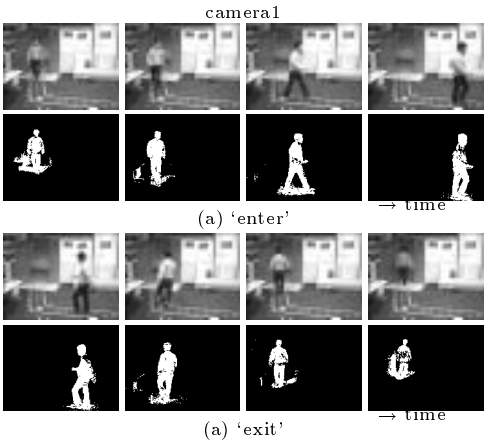

図11　学習用データの例(原画像, 変化領域)
Fig. 11　Examples of training data (gray-level image, anomalous region).

動作認識システムの性能評価を行うためのテストデータは，学習用データとは別に用意した．特に，この実験では複数対象の動作が正しく認識できるか否かを評価するため，複数対象が同時に動作を行う３種類のシーン(２人の「入室」，２人の「退室」，複数人数

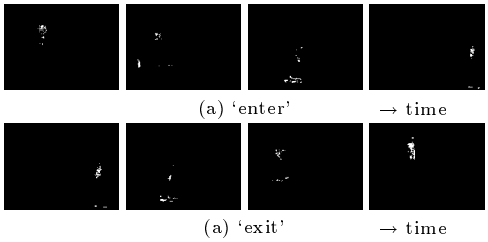

図12　注目領域の一部
Fig. 12　Example of focusing regions.

の「入室」,「退室」) を各々20通り用意し，合計60組 (120本) の映像を撮影して，それをテストデータとした．テストデータの一例 (一部) を図13に示す．

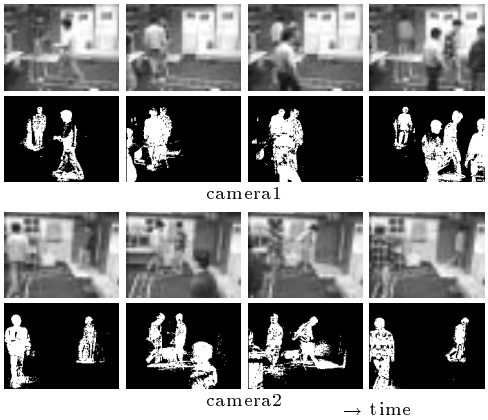

図13　テストデータの例 (一部分)
Fig. 13　An example of test data(extracted).

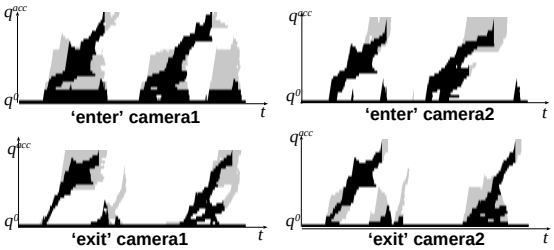

図14　図13の入力に対する状態遷移の例 (黒：活性な状態，灰色：抑制された状態遷移)
Fig. 14　State Transitions for Fig. 13 (black: activated states,gray: inhibited states).

この実験では，２台のカメラで撮影した映像が利用できるので，(a) カメラ１のみを使用，(b) カメラ２のみを使用，(c) 映像レベルの統合法，(d) イベントレベルの統合法，(e) 状態レベルの統合法，の合計５種類の動作認識を行うことができる．

このうち，(e) の手法を用いて図13の映像 (２人が「入室」，２人が「退室」) を認識したときの状態遷移の様子を図14に示す．この図の灰色で示した部分は抑制された状態遷移を表しており，単独の動作認識システムよりも，認識の曖昧性が低減できていることが確認できる．

テスト用の映像60組に対して，上記 (a)〜(e) の５種類の方法で，動作の種別と対象の個数を認識させた結果を図15に示す．この図において，縦軸は正しく認識された映像の個数，横軸はイベント検出に用いた閾価 θ を表している．また，グラフ中の網掛けはテスト映像のシーン種別を表している．

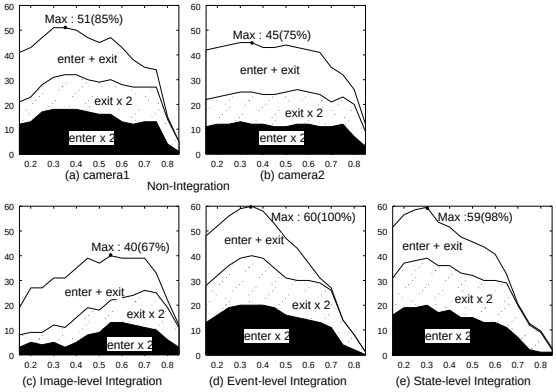

図15　認識結果 (縦軸:正解数,横軸:閾値 θ)
Fig. 15　Recognition results (Vertical:Number of correct recognition,Horizontal: threshold θ).

この実験結果から，以下のことが言える．

まず，全ての認識法において，

（１） θ が大きすぎると状態遷移が途中で途切れやすくなる

（２） θ が小さすぎると複数対象を分離しにくくなるという傾向があり，動作の種別と個数を共に正確に認識するには，最適な θ を与える必要があると言える．

方法によって最適な θ は異なるが，各方法での最適値を比較すると，映像レベルの統合法は単一視点の映像を用いるよりも認識率が低下する．イベントレベル，状態レベルの統合法では，高い認識率が達成されている．と言える．

理論的に最も優れているはずの状態レベルの統合法がイベントレベルの統合法よりもわずかに認識率が低くなっている理由は，学習データの量が少ないため，自由度の高い状態レベルの統合法では十分な汎化が進んでいないためであると考えられる．

5. おわりに

本稿では，複数対象の動作識別と個数の認識を同時

に行う「複数動作の認識問題」という従来の技術では解決が困難であるとされてきた問題を取り挙げ，その一解決法を示した．

まず，選択的注視，対象弁別という２つの機構を持つ単視点映像を用いた複数動作の認識システムを提案した.「選択的注視機構」は，注目領域内でのイベント検出により，注目する動作以外の影響を受けない特徴抽出を行うことができ，状態遷移モデルとしてNFAを採用することにより，複数動作の同時識別が行えるという特徴を持っている．この機構に，NFAの状態遷移とともに動作対象を表す色付きトークンを伝搬させる「対象弁別機構」を導入することにより，複数対象の動作識別とともに個数の認識も同時に行う「複数対象の動作認識システム」を構成する方法を示した．

さらに，より正確な動作認識を行うため，多視点映像を用いた動作認識システムへの拡張法について検討を行った．多視点映像を用いた動作認識を実現するには，映像，イベント，状態，の３つのレベルのいずれかで，多視点で得られた情報を統合する必要があり，そのうち「状態レベルの統合法」が最も能力が高いことを示した．

実験結果から，本論文で提案したシステム，特に多視点映像を用いてイベント，状態の各レベルで情報統合を行うシステムによって，複数動作が正確に認識できることを示した．

本手法は，剛体によって拘束された動作を，対象の位置・形状の時間的変化をもとにして認識する手法であり，動作の空間的変動には対処することはできない．また，背景差分を用いているため，照明条件が変化する場合には正しい認識結果は期待できない．今後は，これらの問題について検討を行い，実環境での使用に耐え得る複数動作の認識システムを構築する予定である．

謝　　辞

本研究の一部は，日本学術振興会未来開拓学術研究推進事業(プロジェクト番号:JSPS-RFTF96P00501)，および文部省科学研究費（基盤研究(A)(2)08408010）の補助のもとに行われたものである．また，初期の試作システム作りに協力して頂いた岡山大学の加藤丈和氏に感謝致します．

文　　献

[1] Yamato J., Ohya J., and Ishii K., "Recognizing human action in time-sequential images using hidden markov model", Proc. of **CVPR**, pp. 379-385, (1992)

[2] Starner T. and Pentland A., "Real-time American sign language recognition from video using hidden markov models", Proc. of **ISCV**, pp. 265-270, 1995.

[3] Bregler C. and Omohundro S.M., "Nonlinear manifold learning for visual speech recognition", Proc. of **ICCV**, pp.494-499, 1995.

[4] Wilson A. and Bobick A., "Learing Visual Behavior for Gesture Analysis", M.I.T. Media Laboratory Perceptual Computing Section Technical Report No.337. 1995.

（平成10年10月20日受付，12月25日再受付）

和田　俊和　（正員）

昭和６２年東工大大学院修士課程修了．平成２年同大大学院博士課程修了．同年岡山大学工学部助手．平成６年同大大学院自然科学研究科助手．平成７年同大工学部講師．平成９年京都大学大学院工学研究科助教授．現在に至る．工学博士．画像理解，パターン認識の研究に従事．平成７年第５回 David Marr 賞．ＩＥＥＥ，情報処理学会，電子情報通信学会会員．

佐藤　正行

平８年京都大学工学部電気電子工学科卒．平１０年同大学大学院電子通信工学専攻修了．現在三菱電機株式会社勤務．

松山　隆司　（正員）

昭５１年京都大学大学院工学研究科修士課程修了．京都大学工学部助手，東北大学工学部助教授，岡山大学工学部教授を経て，平成７年より京都大学大学院工学研究科教授．現在同大学院情報学研究科教授．昭和５７～５９ 米国メリーランド大学客員研究員．工学博士．画像理解，人工知能，分散協調処理に興味を持っている．昭和５５年情報処理学会創立２０周年記念論文賞．平成２年人工知能学会論文賞．平成５年情報処理学会論文賞．平成６年本会論文賞．平成７年第５回 David Marr 賞．平成８年IAPRフェロー．著書「A Structural Analysis of Complex Aerial Photographs」(PLENUM),「SIGMA: A Knowledge-Based Aerial Image Understanding System」(PLENUM),「パターン理解」(オーム社) など．