コンピュータビジョンと
イメージメディア 103－6
（1997．1．23）

選択的注視に基づく動作識別
－分散協調視覚システムにおける対象の動作認識法－

和田 俊和 加藤丈和
岡山大学 工学部 情報工学科
〒700 岡山市 津島中 3－1－1

本稿では、分散協調視覚システムにおける「観測ステーション群による移動対象の協調的監視」を実現するための第1歩として、個々の観測ステーションが持つべき視覚監視機能について述べる。視覚監視タスクにおける動作認識問題では、「複数の動作が同時に観測され得る」ため、通常の動作認識法を用いることはできない。本稿では、この困難さを克服するために、「ある画像上の領域（注目領域）内での画像特徴の検出（『イベント検出』）結果に応じて、注目領域の更新（『アクション起動』）を行なう」という「**選択的注視に基づく動作識別法**」を提案する。本手法では、イベント検出とアクション起動を繰り返す過程で、対象の動作段階に対応する内部状態の遷移を行なうが、非決定性状態遷移モデルを用いることにより、全ての可能な動作状態を生成・追跡することができる。本手法のアクションを「カメラパラメータの更新」、「他の観測ステーションへのメッセージ送信」へと拡張することにより、能動視覚、共同注視の機能を実現することができ、本手法で示す枠組に基づいて、実時間性、頑健性、柔軟性を兼ね備えた分散協調視覚システムを設計することが可能になる。

Motion Recognition by Selective Attention Model for Cooperative Distributed Vision System

Toshikazu Wada and Takekazu Kato
Department of Information Technology, Faculty of Engineering,
OKAYAMA UNIVERSITY
3-1-1 Tsushima-naka, Okayama,700, JAPAN

This paper presents a motion-recognition method for visual-surveillance tasks in Cooperative-Distributed-Vision (CDV) system. This method is designed as a motion recognition mechanism of Observation Station, so as to be capable of the real-time recognition of multiple-object motions. To realize multiple-motion recognition, we propose *Selective Attention Model*, which iterates *event detection* in focusing regions on the image space and the *action* renewing the focusing regions according to the event detection result. While iterating the event detection and action, the model changes its internal states corresponding to the object-motion stage. By employing an NFA as its state transition mechanism, the model can generate and follow all possible states. The *action* of this model can be expanded as "camera-parameter control" and "message sending to other stations", which enables *active* and *cooperative* motion recognition.

1 はじめに

カメラから入力される動画像を解析することによって、シーン中を移動する対象の振舞いを認識・理解する視覚監視の問題は、分散協調視覚による広域分散監視 [1] を実現する上で解決すべき最も基本的な問題の一つである。

実環境で安定かつ正確に動作する視覚監視システムを構成するためには、まずシーンの撮影法について検討する必要がある。視覚監視のための撮影法としては、対象を画像枠にかからないようにできるだけ大きく撮影し、その適切な撮影状態を維持しながら対象をできるだけ長時間観測し続けることが望ましい。このようなカメラワークを伴う撮影法を用いた場合の画像解析は一般に困難であるが、我々は運動視差が生じないように調整したパン-チルト-ズームの変更が可能な可動カメラを用いて撮影し、撮影された画像に対して背景差分を行なう方法 [2] を提案している。

しかし、背景差分によって得られる情報は、単に背景とは見え方が異なる「変化領域」の情報であるので、対象だけでなく、対象の移動に付随する他の物体の移動、対象の影や反射などによっても変化領域が生じるため、変化領域自体を対象の投影像と見なすことはできない。この情報からいかにしてシーン中の「対象」に関する情報を獲得するかが重要な問題となる。

これを実現する方式としては、

- 画像入力、通信、計算の機能を持つ「観測ステーション」を複数台用いて得られた背景差分情報から、各時刻で3次元シーン中の対象の個数、位置、物理的形状を求める方式。
- 対象の動作パターンが分類可能な場合、観測された動画像から対象の動作を識別する方式。

の2つが考えられる。

前者の方式は、対象の動きに関する制約条件が不要であり適用範囲も広いが、制約条件が緩いため、信頼性の高い結果を得るには、視野を共有する複数台の観測ステーションが協調的に処理を行なう必要が生じる。

一方、後者の方式は、対象がシーン中の壁や通路、ドアなどの剛体に拘束された動作をする場合に適用可能であり、対象の動作に関する事前知識を利用することができるため、1台の観測ステーションだけでも比較的安定な認識を行なうことができるものと考えられる。また、視野を共有する複数台の観測ステーションから得られた情報を利用することができれば、より信頼性の高い結果が得られるものと期待される。

したがって、広場など対象の動きを拘束する物体が存在しない状況下では前者の方式を採用し、屋内や通路などでは後者の方式を採用すれば良いと言える。

本研究では、観測ステーションで得られた背景差分の結果から、シーン中での対象の動作を識別する方法について検討を行なう。

1.1 視覚監視における動作認識問題

視覚監視における動作認識問題は、ジェスチャー認識など通常の動作認識問題とは異なり、

- 1つの画像フレームに複数の認識対象が現れ得る。

という特殊性がある。これは、視覚監視タスクでは、観測された動画像を空間的に分割して扱わなければならないことを意味している。

これまでに提案されている画像による動作認識法の多くは、画像から点や線、領域などの画像特徴を抽出しないで画像全体を単なる入力パターンと見なす「見え方 (appearance) に基づく手法」と HMM などの状態遷移モデルを組み合わせた手法 [3],[4],[6] である。これらの手法は、ジェスチャー認識や、手話認識など、単一の動作だけが観測されるタスクには向いているが、本研究で扱う「複数の対象の動作を同時に認識する問題」には適用することはできない。一方、入力画像からボトムアップ処理のみによって画像特徴を抽出する手法 (bottom-up feature extraction) は、動作系列を解析するための状態遷移モデルと組み合わせた場合、各画像に対する特徴抽出の誤りが蓄積されてしまうためシステム全体が不安定になりやすく、画像特徴が比較的抽出しやすい lip reading[5] などの限られた用途でしか用いられていないのが現状である。

また、通常の識別問題は、事象の全体集合が与えられたとき入力をその集合に含まれる1つの要素に対応付ける問題であるので、識別候補の間に排他性が成り立つものと仮定し、ある識別結果を支持する Positive な証拠情報と、それを支持しない Negative な証拠情報の両方が利用できる。ところが、本稿で取り扱う問題では、全事象として動作集合の巾集合を考えなければならないため、1）個々の動作に関する Negative な証拠情報を用いることができない、2）複数の識別結果が考えられた場合、それらが曖昧さを表しているのか、同時に複数の動作が起きたことを表しているのかの区別がつかない、という認識問題としての困難さが伴う。

1.2 提案手法の概要

以上のような既存の動作認識法の問題点を考慮し、本研究では、動作シーケンスの解析結果を利用して画像特徴の抽出をトップダウン的に行なう「選択的注視」機構を提案し、これによって与えられた動画像から複数動作を同時に認識する手法について述べる。本研究で提案する動作認識法は、利用する特徴に応じて様々な拡張が可能であるが、本稿では背景差分によって得られる変化領域の特徴を利用する場合について述べる。

本研究では、動画像データのクラスに固有の特徴が現れる画像中の場所（「注目領域」）の系列が存在することを仮定し[1]、この注目領域内部で、特定の画像特徴が抽出できたか否かに基づいて動作認識を行なう手法を提案する。

[1] ここで導入された仮定が妥当性を持つのは、シーン中に壁や机、ドアなど、固定、あるいは関節が固定された剛体が存在し、シーン中を移動する物体が、これらの剛体に拘束された動きをする場合である。このような剛体が監視の対象となるシーン中に含まれる場合は多く、導入された仮説によって本手法の適用範囲が著しく狭められることはない。

例えば、人がドアを開けて部屋から出ていく動画像では、まずドアノブ部分の画素値が変化し、次にドアの縁の画素値が変化し、… という具合に、動作に固有の画像特徴が現れる場所が決まっている。この場合、ある時刻に画素値の変化が起きた部分から、次に変化が起きる部分を容易に予測することが可能であり、注目領域を更新しながら動作認識を行なうことができる。以上のように、注目領域を動的に更新しながら、その内部で特徴抽出を行なう手法を「**選択的注視**」と呼ぶ。

選択的注視に基づいて、あるクラスの動作を撮影した動画像を受理する**動作同定機構**を構成することができる。同定機構は、

- 各動画像クラスの注目領域シーケンス
- 注目領域内部での画像特徴の有無を調べるイベント検出器
- 動作を記述する状態遷移モデル

から構成される。イベント検出器は、観測画像について現在の状態に対応する注目領域内部で画像特徴が検出できたか否かの「イベント情報」を状態遷移モデルに送り、状態遷移モデルは現在の状態とイベント情報に応じて状態遷移を起こすという手続きを繰り返す。この手続きは、イベント情報を検出して状態遷移を行なうボトムアップ処理と、現在の状態に応じて注目領域を更新するトップダウン処理を有機的に結合したものであり、これによって注目領域を動的に更新しながら状態遷移を計算することができる。

状態遷移モデルとして、非決定性有限オートマトン (NFA) を導入することにより、各時刻において複数の状態を持つことが許容されるようになる。この結果、

- 入力に対して考えられる全ての状態を保持することにより、画像特徴のシーケンスを見失うことなく安定に辿ることができる。
- 同時に撮影された複数の対象の動きを個別に認識することが可能である。

という特長が得られる。

複数クラスの動作識別機構は、同定機構を複数用いて構成されており、各同定機構の注目領域シーケンスは、他のクラスとは独立に学習可能である。このため、個別にトレーニングされた複数の同定機構を組み合わせて、多クラスの動作識別機構を容易に構成することができるという利点がある。

前述のように、動作識別機構は同時に複数の状態を保持するため、動作識別を行なうには、ある時刻に得られた複数の状態から「動作に関する仮説」を生成しなければならない。これを「状態集合の解釈」と呼ぶ。状態集合の解釈が必要になる理由は、

- 複数の要素から成る状態集合に対して 1）各状態に対応する動作が同時に起きている、2）各状態に対応する動作のいずれであるかが曖昧である、というように複数の解釈が考えられるため、この多義性を必要最小限に押える。

表 1: 長さ 2 のイベントコードに対する状態遷移表 (現在の状態：q^k)

$\sigma^i = e(f(q^k), I^i) \cdot e(f(q^{k+1}), I^i)$	$\delta(q^k, \sigma^i)$
$0 \cdot 0$	q^{rej}
$0 \cdot 1$	q^{k+1}
$1 \cdot 0$	q^k
$1 \cdot 1$	q^k, q^{k+1}

表 2: 長さ 3 のイベントコードに対する状態遷移表 (現在の状態：q^k)

$\sigma^i = e(f(q^k), I^i) \cdot e(f(q^{k+1}), I^i) \cdot e(f(q^{k+2}), I^i)$	$\delta(q^k, \sigma^i)$
$0 \cdot 0 \cdot 0$	q^{rej}
$0 \cdot 0 \cdot 1$	q^{k+2}
$0 \cdot 1 \cdot 0$	q^{k+1}
$0 \cdot 1 \cdot 1$	q^{k+1}, q^{k+2}
$1 \cdot 0 \cdot 0$	q^k
$1 \cdot 0 \cdot 1$	q^k, q^{k+2}
$1 \cdot 1 \cdot 0$	q^k, q^{k+1}
$1 \cdot 1 \cdot 1$	$q^k q^{k+1}; q^{k+2}$

- 解釈の結果、動作組を要素とする可能な事象の集合が得られるため、可能な動作クラスの全てが既知である場合には、各動作組に対する Negative な情報を利用した、より詳細な識別を行なうことができる。

の 2 つである。本研究では、各状態に対応する変化が起こり得る領域（「可能変化領域」）を導入し、注目領域と可能変化領域の関係によって、状態集合の解釈を行なう手法を提案する。

以下、選択的注視に基づく動作同定機構、識別機構について述べる。

2 選択的注視に基づく動作同定機構

ここで述べる動作同定機構とは、与えられた動画像データが、あるクラスに属するか属さないかを判定する 2 クラスの識別機構である。

選択的注視に基づく動作同定機構は、1）動画像クラスに固有の注目領域シーケンス、2）注目領域内で画像特徴が検出できたか否かを調べるイベント検出器、3）各クラスの動作を記述する非決定性の状態遷移モデル (NFA)、の 3 つから成る。各々の定義は、以下の通りである。

定義 1 (注目領域) 注目領域のシーケンス情報は、状態から画像領域への写像 $f(q)$: $Q \mapsto \mathcal{B}(X \times Y)$ によって表現される。(但し、Q は状態の集合、$X \times Y$ は画素の集合、$\mathcal{B}(A)$ は 集合 A の巾集合を表す。)

定義 2 (イベント検出器) イベント検出器は、画像 I と注目領域 f から、$\{0,1\}$ への写像を与える述語 $e(f, I)$: $\mathcal{B}(X \times Y) \times \mathcal{I} \mapsto \{0,1\}$ を、複数の注目領域につい

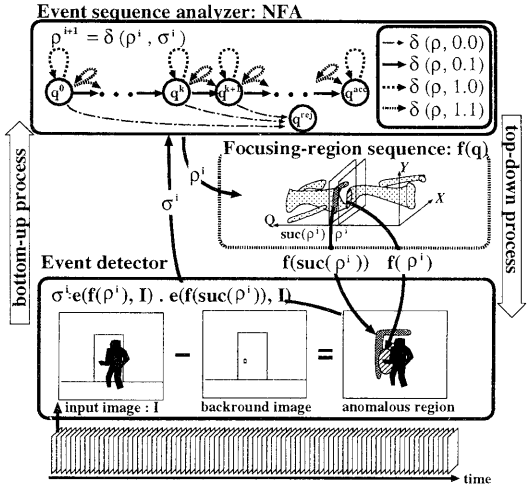


図 1: 動作同定機構の動作（背景差分を用いる場合）

て求め、それらを結合したイベントコードを出力する。但し、$\mathcal{I}$ は画像の集合を表し、$e(\phi, I) = 1$ とする。現在の状態が q^i であるとき、長さ n のイベントコードは、$e(f(q^i), I) \cdot e(f(q^{i+1}), I) \cdots e(f(q^{i+n-1}), I)$ と表される。

定義 3 (動作記述用オートマトン) 動画像の系列は、初期状態：q^0、状態の有限集合:Q、入力記号: Σ、状態遷移関数: δ、最終状態の集合:Q_f の 5 つ組から成るオートマトン $M = (q^0, Q, \Sigma, \delta, Q_f)$ によって表される。但し、Σ はイベントコードの値域、δ は 図 1や 2 に示す状態遷移表で定義され、$Q_f = \{q^{acc}, q^{rej}\}$ （q^{acc} は、$q^0, q^1, \cdots, q^m$ という状態系列の最後 q^m 、q^{rej} は画像系列が受理できないことを表す。）とする。

イベントコードの長さが 2 の場合の動作同定機構の動作を以下に示す（図 1参照）。

1. 現在の状態ρ^0を初期状態 q^0、$i = 0$ とする。
2. $\rho^i = q^{acc}$ であれば、現在までの動画像系列を受理し、結果を出力する。
3. イベント検出器は、イベント・コード$\sigma^i = e(f(\rho^i), I) \cdot e(f(suc(\rho^i)), I)$ を計算し、状態遷移モデルの入力に与える。[a]
4. 状態遷移モデルでは、この入力と現在の状態から次の状態$\rho^{i+1} = \delta(\rho^i, \sigma^i)$ を計算する。
5. $i = i + 1$ とし、2 に戻る。

[a] $suc(\rho^i)$ は後継者 (successor) 関数を表す。

前述の説明では、簡単のため、現在の状態ρ^iが単一の状態のみを表すかのように説明したが、実際には表 1,2に示すように、δ は複数の 1 を含むイベントコードに対して、複数の状態への非決定性の状態遷移を起こすため、各時刻の状態は集合になる。この集合に属する各状態について上述の計算を行なうことになる。

2.1 変化領域からの注目領域の学習

ここでは、背景差分によって得られる変化領域の情報から、注目領域のシーケンスを学習する方法について述べる。

入力画像 $I(x, y, t)$ の変化領域は、背景のみが映された画像 $I_{ref}(x, y, t)$ と閾値 T によって、以下のように求めることができる。

$$a(t) = \{(x, y) | \ |I(x, y, t) - I_{ref}(x, y, t)| > T\} \quad (1)$$

同一のクラスに属する動画像データの変化領域 $a_i(t)$ $(i = 1, 2, \cdots, n)$ が与えられた時、まず、各 $a_i(t)$ の時間軸を揃えるため、次式の値を最大化する時間軸の非線形伸縮関数 τ_i を求める。

$$\int \frac{|a(t) \cap a_i(\tau_i(t))|}{|a(t) \cup a_i(\tau_i(t))|} dt \quad (2)$$

但し、τ_iは単調関数、$a(t)$ は $a_i(t)$ の中から選ばれた標準データ、$|\cdot|$ は、画素数を与える演算子とする。この計算は動的計画法 (DP) を用いることにより容易に計算することができる。

時間軸が正規化された変化領域データから、注目領域 $f(t)$ は以下のように計算することができる。

$$f(t) = \bigcap_{i=1}^{n} a_i(\tau_i(t)). \quad (3)$$

以上の計算法は、以下の逐次的計算に置き換えることができる。

$$f_1(t) = a_1(t),$$

$$f_i(t) = f_{i-1}(t) \cap a_i(\tau_i(t))$$

但し、ここでは $a_1(t)$ を標準データ $a(t)$ としている。

この計算においては、変化領域のデータ a_i $(i > 1)$ は f_i を計算した後に削除することができるため、大量のデータを用いた学習が可能となる。

以上の計算によって得られた注目領域 $f(t)$ は、標準データ $a(t)$ の時間軸で表現されており、この時間軸を適当な間隔に区切ったものが、状態 q となる。各状態の注目領域は、状態に対応する時間間隔中の $f(t)$ の集合積によって表される。すなわち、状態 q に対応する時間間隔が $t_s \leq t < t_e$ であったとすると、状態 q に対応する注目領域 $f(q)$ は、以下のようになる。

$$f(q) = \bigcap_{t=t_s}^{t_e - 1} f(t) \quad (4)$$

2.2 背景差分を用いた場合のイベント検出

背景差分によって得られる変化領域の情報を用いた場合、イベント検出結果 $e(f, I)$ を以下のように表現することができる。

$$e(f, I_i(t)) = \begin{cases} 1, & f \cap a_i(t) = f \\ 0, & otherwise \end{cases} . \quad (5)$$

これは先の学習アルゴリズムと併用した場合、学習サンプルに対しては常に、$e(f(t), I_i(\tau_i(t))) = 1$ となる。しかし、実際には学習サンプルと入力データとは完全に一致しないので、イベント検出結果は以下のように定義すべきである。

$$e(f, I_i(t)) = \begin{cases} 1, & \frac{|f \cap a_i(t)|}{|f|} > \theta \ or \ f = \phi \\ 0, & otherwise \end{cases} \quad (6)$$

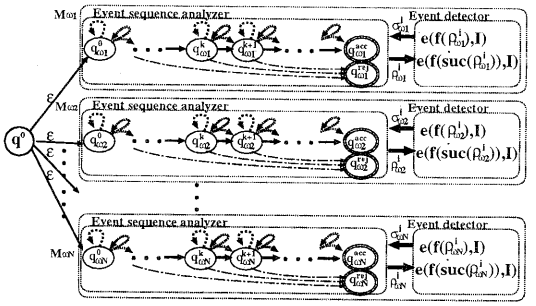


図 2: 動画像識別機構

但し、θ $(0 < \theta \leq 1)$ は閾値である。

式6は、注目領域内部で変化画素の占有率が閾値θを越えた場合に、イベント検出結果が1となり、注目領域が空 $(f = \phi)$ の場合には任意の画像に対してイベント検出結果は常に1となることを表している。

この場合、初期状態における注目領域 $f(q^0)$ を空集合ϕ とすることによって、$e(f(q^0), I) = 1$ となり、常に現在の状態の集合に初期状態が含まれるようになる。これによって、処理を中断したり後戻りすることなく、入力され続ける動画像を継続的に識別することができるようになる。このため、前述の手続きには停止条件を含めていない。

3 動作識別機構

動画像識別機構はクラス ωi $(i = 1, \cdots, N)$ の同定機構 $M_{\omega i} = \{q^0_{\omega i}, Q_{\omega i}, \delta_{\omega i}, e, f_{\omega i}\}$ から構成される。具体的には、新たな初期状態 q^0 と、q^0 から各同定機構の初期状態 $q^0_{\omega i}$ へのϵ-遷移[2]を付け加えることにより、図2に示す識別機構を構成することができる。

3.1 状態集合の解釈

識別機構は同定機構と同様に非決定性の状態遷移を含んでおり、各時刻で複数の状態 $\{\rho^k\}$ を持ち得る。したがって、様々な状態の組合せから、どのような動作が起きていると判定すべきかという「状態集合の解釈」に関するルールを定めなければならない。

まず、単一の同定機構 $M_{\omega i}$ が持つ状態の組合せについて考えてみる。

- 状態 $q^{rej}_{\omega i}$ と $q^0_{\omega i}$ [3]のみが存在する場合、同定不能であると解釈できる。
- 状態 $q^k_{\omega i}$ $(k \neq 0 \ \wedge \ k \neq rej)$ が存在する場合、
  1. 他に $q^0_{\omega i}$ あるいは $q^{rej}_{\omega i}$ [4]以外の状態が存在しない場合、$q^k_{\omega i}$ に対応する動作が起きていると判定する。
  2. 他に状態 $q^m_{\omega i}$ $(m \neq 0 \ \wedge \ m \neq rej \ \wedge \ m \neq k)$ が存在する場合には、1)「$q^k_{\omega i}$ と $q^m_{\omega i}$ に対応する動作が同時に起きている」2)「$q^k_{\omega i}$ と $q^m_{\omega i}$ に対応する動作のいずれが起きているのかが不明確である」という2種類の解釈ができる。

次に、異なる同定機構の間での状態の組合せについて考えてみる。

- 全てのクラスの同定機構が、$q^0_{\omega j}$ あるいは $q^{rej}_{\omega j}$ の状態しか持たない場合は、識別不能であると解釈できる。
- あるクラス ωi の同定機構が状態 $q^k_{\omega i}$ $(k \neq 0 \wedge k \neq rej)$ を持つとき、他のクラスの状態について次の2つの場合が考えられる：
  1. 他のクラスωj $(j \neq i)$ の状態が $q^0_{\omega j}$ あるいは、$q^0_{\omega j}$ と $q^{rej}_{\omega j}$ のみである場合、$q^k_{\omega i}$ に対応する動作が起きていると判定する。
  2. 他のクラスωj $(j \neq i)$ の状態が、$q^m_{\omega j}$ $(m \neq 0 \ \wedge \ m \neq rej)$ を含む場合には、1)「$q^k_{\omega i}$ と $q^m_{\omega j}$ に対応する動作が同時に起きている」2)「$q^k_{\omega i}$ と $q^m_{\omega j}$ に対応する動作のいずれが起きているのかが不明確である」という2種類の解釈ができる。

以上を整理すると、$q^k_{\omega i}$ $(k \neq 0 \ \wedge \ k \neq rej)$ という状態が同時に複数現れる時、状態集合の解釈に多義性が生じることが分かる。

3.1.1 可能変化領域による状態集合の解釈と否定情報の取り扱い

ここで、解釈を以下のように表現するものとする。

- “状態1” に対応する動作が起きているという解釈を、“状態1” で表す。
- “解釈1” と “解釈2” に対応する動作が同時に起きているという解釈を “解釈1 and 解釈2” と表す。
- “解釈1” もしくは “解釈2” であるという解釈を、“解釈1 or 解釈2” と表す。
- 解釈を表現する場合、曖昧さが生じないように適宜かっこを用いる。

また、状態集合に対する解釈を与えるために、以下に述べる可能変化領域を導入する：

これまでに用いてきた集合積によって表される注目領域 $f(\cdot)$ は、引数に対応する状態において全学習サンプルの画素値の変化が必ず観測される画像上の領域を表しており、「**共通変化領域**」と呼ぶことができる。同様に、ある状態において変化する可能性がある「**可能変化領域**」を集合和によって求めることができる。これを $F(\cdot)$ で表す。

ここで、状態 A と B が同時に存在する場合を考えると、明らかに $A$and$B$ という解釈が成立する。これ以外の解釈があり得るかどうかを $F(A)$ と $f(B)$、$f(A)$ と $F(B)$ の関係に基づいて整理すると、以下のようになる。

- $f(A) \not\subseteq F(B) \ \wedge \ F(A) \supseteq f(B)$ の場合 (図3 (a)): A に対応するデータが与えられただけで、状態 B が生じてしまったと考えることもできるので、解釈 “A” が成り立つ。全体の解釈は $(A\text{and}B)\text{or}A$ となる。

[2] ϵ-遷移とは空入力に対する状態遷移のことである。
[3] 状態 $q^0_{\omega i}$ は常に存在するので、解釈の際には無視する。
[4] この場合、$q^{rej}_{\omega i}$ は誤ったシーケンスの追跡の結果生じたものであると考えられるため、無視する。

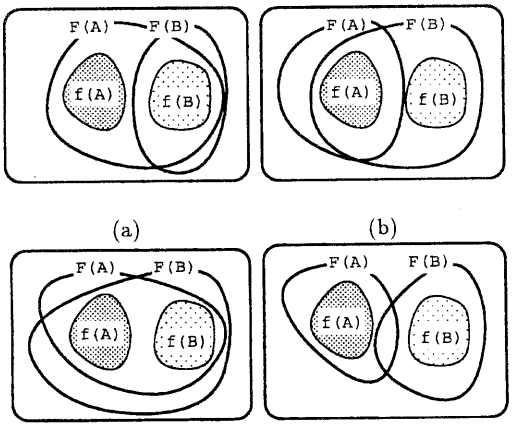


図 3: $F(A)$ と $f(B)$、$f(A)$ と $F(B)$ の関係の分類

- $f(A) \subseteq F(B) \wedge F(A) \not\supseteq f(B)$ の場合 (図 3 (b)): B に対応するデータが与えられただけで、状態 A が生じてしまったと考えることもできるので、解釈 "B" が成り立つ。全体の解釈は $(A\text{and}B)\text{or}B$ となる。
- $f(A) \subseteq F(B) \wedge F(A) \supseteq f(B)$ の場合 (図 3(c)): A に対応するデータが与えられただけで、状態 B が生じてしまった、あるいは B に対応するデータが与えられただけで、状態 A が生じてしまったと考えることもできるので、解釈 "A" および 解釈 "B" が成り立つ。全体の解釈は $(A\text{and}B)\text{or}A\text{or}B$ となる。
- $f(A) \not\subseteq F(B) \wedge F(A) \not\supseteq f(B)$ の場合 (図 3(d)): この場合、他の解釈は存在し得ない。全体の解釈は $(A\text{and}B)$ となり、曖昧さは生じない。

但し、関係 $f \subseteq F$ は、次式が満足される時に成立するものとする。

$$\frac{|f \cap F|}{|f|} > \theta \tag{7}$$

解釈を生成する際には、あらかじめ任意の状態の組 A, B に関して関係 $f(A) \subseteq F(B)$ を調べて表にしておき、複数の状態から成る状態集合が得られたとき、この表を参照しながら解釈を動的に生成する。

以上のように、可能性として考えられ得る全ての解釈を生成することにより、解釈の全体集合が明らかになり、or でつながれた解釈のうち、いずれかが正しいことになる。

解釈の生成は、状態 $q_{\omega i}^{acc}$が現れた時点で行なうものとし、各時刻での解釈生成は行なわない。これは、解釈のシーケンスの矛盾が生じた場合、それを正当化する高次の解釈が常に存在するため、解釈シーケンスの解析によって可能な解釈集合が縮小しないためである。

否定情報の利用は、得られた解釈に含まれる or で継れた各動作組に対応する可能変化領域の集合和を求め、それ以外の部分で検出された変化領域が観測された変化領域に対して与えられた閾値θ' 以上現れた時、その動作組を解釈から取り除くという操作で実現できる。

3.1.2 関係の学習

$f(A) \subseteq F(B)$ であるとき、実際に状態 B に対応する画像の変化領域が $f(A)$ を覆うかどうかは明らかではない。なぜならば、単に $F(B)$ を求める際の集合和の演算によって $F(B)$ が膨張したため $f(A) \subseteq F(B)$ となってしまっただけで、実際には B に対応する動画像フレームのインスタンスは $f(A)$ を覆わないかもしれないからである。すなわち、上述の解釈は $f(A) \subseteq F(B)$ であるならば、状態 B に対応する動作のみが起きている「可能性がある」ことを表しているに過ぎない。

これを調べるには、一旦学習した注目領域 f に関して、$f(A) \subseteq F(B)$ を満足する状態 B に対応する画像の変化領域データが実際に $f(A)$ を覆うかどうかを調べれば良い。実際には、$f(A)$ を覆う状態 Bに対応するデータが存在しなければ、関係 $f(A) \subseteq F(B)$ を表から取り除き、解釈 B が導かれないようにする。このような学習を「関係の学習」と呼ぶ。

関係の学習によって、必要以上に解釈の個数が増加するのを防ぐことができるが、一方で学習の簡便さが損なわれるという問題も生じる。このような利害得失と、実際のタスクの持つ性質を考慮して関係の学習を行なうか否かを決めなければならない。

3.1.3 解釈の候補と遅延評価

$q_{\omega i}^{acc}$ が現れた時点で同時に存在する他の状態のうち、$q_{\omega j}^{acc}$ $(i \neq j)$ は確定しているが、$q_{\omega j}^{m}$ $(m \neq 0, m \neq acc, m \neq rej)$ は最終的に $q_{\omega j}^{acc}$に到達するか否かは決まっていない。その意味で、$q_{\omega i}^{acc}$ が現れた時点で生成した解釈はあくまで解釈の候補に過ぎない。

したがって、$q_{\omega i}^{acc}$ が現れた時点で同時に存在する他の状態のうち、$q_{\omega j}^{m}$ $(m \neq 0, m \neq acc, m \neq rej)$ に関しては全てマークをつけておき、それらが $q_{\omega i}^{acc}$ に到達した場合あるいは $q_{\omega i}^{acc}$ に到達し得ないことが判明した段階で、過去に下した解釈が確定する。

具体的には、$q_{\omega i}^{acc}$ が現れた時点で同時に存在する他の状態が

- $q_{\omega i}^{acc}$ に到達した場合には、解釈に含まれる対応部分を確定させる。
- $q_{\omega i}^{acc}$ に到達し得ないことが判明した場合には、対応する状態を取り除いて解釈を作り直す。

という操作を行なう。

この結果、$q_{\omega i}^{acc}$ が現れた時点でどのような対象の動作が起きていたのかという解釈が確定し、確定した解釈間に矛盾が生じることはない。

確定した解釈に含まれる or で継れた個々の解釈のうち、いずれかは正しいので、別の評価尺度を導入することにより解釈を 1 つに絞り込むことが可能になる。これは、共通の動作を学習させた視野を共有する複数のシステムが生成した解釈の論理積によって、可能な解釈を絞り込むなどの応用が可能であることを示唆している。

4 実験

ここでは、以上に述べた動画像識別システムを試作し、ドアを開けて人が部屋に出入りする様子を撮影した動画像の識別に適用した結果を示す。

動画像のクラスは、「入室」と「退室」の2つとし、各々8個、合計16個の動画像データを用意した。画像フレームのサイズは 128×120 とし、30[フレーム/秒] でサンプリングを行なった。各動画像データを構成するフレーム数は 167～319 である。

「入室」のクラスに属する動画像データには、入室後人が左の画像枠に向かって移動するデータと右の画像枠に向かって移動するデータの2種類が含まれている。また、「退室」のクラスの動画像データも人が左と右からドアに向かって移動するデータが含まれている。これらの動画像データに対する背景差分を適用し、閾値20で変化領域の検出を行なった。各クラスの動画像データの例と、それらの背景差分結果を図4,5に示す。

また、退室、入室それぞれのクラスの全背景差分データに対して、DPマッチングを行ない注目領域と、可能変化領域を求めた結果を図6~10 に示す。但し、図6は時間軸方向の量子化を行なう以前の注目領域であり、図7~10は、時間軸の正規化の後、10フレーム (1/3秒) の均一な時間間隔でデータを区切り、同じ時間間隔内のデータに対する AND と OR の集合演算で求めている。

識別実験は、16個のデータのうちの1つを入力、残り全てを学習用として、イベント検出に用いる閾値θを $1 \sim 0.9$ の範囲で変化させながら識別する実験を、入力データを入れ換えて全データについて行なった。但し、イベントコード長は3であり、関係の学習は行なっておらず、否定情報による解釈の取り消しに用いる閾値は $\theta' = 1 - \theta$ とした。

識別結果を図11に示す。横軸が閾値 θ、縦軸がデータの個数である。解釈は一般には複数個得られて繁雑であるので、この図では識別結果を以下の4つに分類している。

正解 (A) 解釈に含まれる状態が全て入力データのクラスに対応している場合

曖昧 (B) 解釈に含まれる状態が正解を含む複数のクラスに対応している場合

誤識別 (C) 解釈に含まれる状態が全て誤っている場合

棄却 (D) 最終状態に到達しなかった場合

閾値が 0.976～0.962 の範囲で、16個のデータのうち14個が正しく識別されており、0.965 以下で誤識別が無くなっている。

さらに、全学習データから注目領域、可能変化領域のデータを作成し、図12に示す3人の動作（1人が退室、1人が入室、1人がドアの前を通り過ぎる）が撮影された動画像データに対する解釈を閾値 0.99 で求めたところ、125, 127, 249 番目の画像で生成された解釈が、それぞれ 144,144,280 番目の画像が入った時点で $(Enter)$, $(Enter \text{ and } Enter)$ or $(Enter)$, $(Exit)$ と確定し、正しい解釈が含まれることが確認できた。

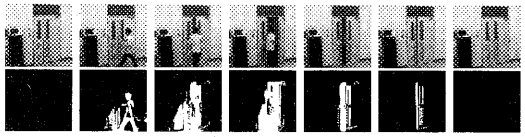


図 4: “退室” の動画像と背景差分の例. ⟶ time

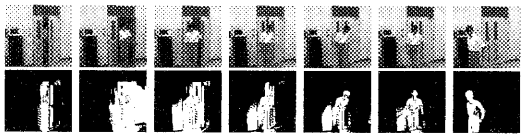
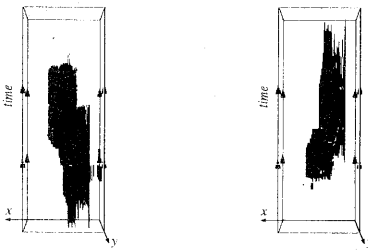


図 5: “入室” の動画像と背景差分の例. ⟶ time



(“入室”) (“退室”)

図 6: 時間軸方向の量子化以前の注目領域

図 7: “退室” の注目領域 → state

図 8: “退室” の可能変化領域 → state

図 9: “入室” の注目領域 → state

図 10: “入室” の可能変化領域 → state

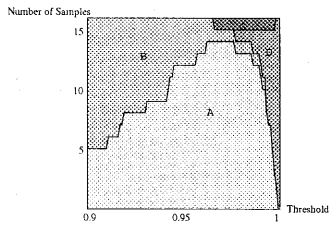


図 11: 識別結果 (A:正解,B:曖昧,C:誤識別,D:棄却)

No.60 (1) No.90 (3) No.105 (3) No.120 (3) No.130 (3) No.150 (2) No.180 (1)

図 12: 複数の動作が含まれるデータ (括弧内は対象の個数)

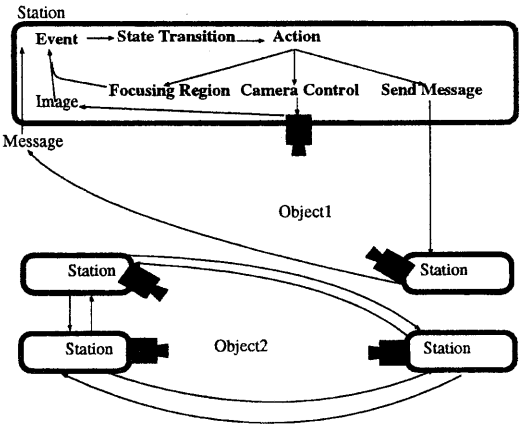


図 13: 共同注視による動作認識

5 まとめと今後の課題

本報告では、複数の動作を行なう対象が同時に観測され得る状況下で、背景差分によって得られる画像の変化領域の情報を用いた動作認識法を提案した。本手法には以下のような特長がある。

- 「選択的注視」というトップダウン的な特徴抽出機構を用いることにより、複数の動作を連続的に認識することができる。
- 注目領域内部で簡単な画像処理を行なっているだけなので、高速な動作認識が実現できる。
- 注目領域の学習は、個々の同定機構で独立に行なうことができるので、多クラスの動作認識システムを容易に構成することができる。

本研究では、複数の対象が同時に観測され得る状況を想定しているため、個々の動作に関しては、注目領域（共通変化領域）内で観測される Positive な情報を用いた解析しか行なうことができない。しかし、全ての可能な動作クラスに関する学習ができている場合、「状態集合の解釈」を行なうことにより、解釈に含まれる "or" で継れた個々の動作組に関しては否定情報を利用できるようになる。この否定情報の利用と、遅延評価、関係の学習の 3 つにより、解釈に含まれる曖昧さを抑制することができるようになる。これらの曖昧性を抑制する手法は確かに有効であるが、誤認識が生じないようにすると、多数の解釈が生じることが実験結果からも明らかであり、より正確な動作認識のためには他の画像特徴の利用や、視野を共有する他の観測ステーションとの「共同注視」が必要になると言える。

5.1 今後の課題

ここまでの議論では、画像上での注目領域の設定という仮想的な "アクション" のみを考えてきたが、これは実際のカメラアクションへと拡張することが可能である。すなわち、Appearance Sphere 上に注目領域を拡張して各状態に、1）動作を観測するための最適なカメラパラメータと 2）Appearance Sphere 上の注目領域、を対応付けておけば、画像上で観測されるイベントに応じたカメラアクションと画像上での注視を同時に行なう「能動視覚システム」を構成することができる。このシステムによって、認識対象の動作をできるだけ正確にとらえるようなカメラコントロールが実現でき、動作認識の信頼性も向上するものと考えられる。

さらに、単一の観測ステーションだけでなく、複数の観測ステーションの連携により、より信頼性の高い動作認識を行なう方法も考えられる。すなわち、ある共通の動作を学習した観測ステーションが複数存在する場合、1）ある観測ステーションでその動作クラスに対応する状態への遷移が起きると、他のステーションにそれを通知し、2）通知を受けた他の観測ステーションは、現在処理中のタスクがなければ、その動作を認識するためのカメラアクションを行なうとともに、互いの状態を参照しながら共同で認識処理を行なうのである。このように、イベント-アクションループを複数の観測ステーションに拡張することにより、シーン中の対象を多角的に観測し、より信頼性の高い動作認識を行なう「共同注視」の機能を実現することができる（図 13）。

本研究では、ここで述べた選択的注視の考えに基づいて、1）個々の観測ステーションでの動作認識、2）カメラ・アクションを伴う能動視覚、3）複数の観測ステーションの連携による共同注視の 3 つの機能を実現するための理論的検討、アルゴリズムの開発とシステムの実装を行なう予定である。

参考文献

[1] 和田, 田村, 松山: "広域分散監視システムにおける分散協調型対象同定法", 画像の認識・理解シンポジウム (MIRU'96) 講演論文集, Vol. I, pp. 103-108, (1996)

[2] 和田, 浮田, 松山:"Appearance Sphere – パン・チルト・ズームカメラのための背景モデル –", 画像の認識・理解シンポジウム (MIRU'96) 講演論文集,Vol. II, 103-108, (1996)

[3] Yamamoto J., Ohya J., and Ishii K., "Recognizing human action in time-sequential images using hidden markov model", Proc. of CVPR, pp. 664-665, (1992)

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

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

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