コンピュータビジョンと
イメージメディア 109－13
（1998. 1. 23）

多視点映像を用いた協調的動作認識

佐藤 正行　和田 俊和　松山 隆司
京都大学大学院工学研究科 電子通信工学専攻
京都市 左京区 吉田本町

我々は、単一視点から観測した映像による動作認識手法として、非決定性オートマトン (NFA) とその内部状態に対応する注目領域内での画像の変化 (イベント) の検出を用いた「選択的注視に基づく動作識別」を提案している。この手法では、動作対象を明示的に取り扱っていないため、1つの対象の動作によって時間的に連続した複数の状態への遷移が同時に起き、対象の個数が本質的に認識できないという問題点がある。本稿ではまず、対象概念を導入するため、トークンの伝搬により、動作対象に対応する時間的に連続した状態の部分系列を求めるアルゴリズムを提案する。さらに、認識能力の向上を目的とする 多視点映像を用いた動作認識への拡張の方法として、映像・イベント・状態 の各レベルで多視点映像情報を統合する 3 種類の方式について検討し、異なる NFA の内部状態をそれらの共起性に基づいて統合する手法が一般性・拡張性の点で最も優れていることを明らかにする。これらの「対象概念の導入」「多視点映像への拡張」により、正確かつ安定な認識機構が構成できることを実験によって示す。

Cooperative Behavior Recognition from Multi-Viewpoint Images

M. Sato　T. Wada　T. Matsuyama
Department of Electronics and Communication,
Graduate School of Engineering, KYOTO UNIVERSITY
Yoshidamototyo, Kyoto, JAPAN

We have proposed a behavior recognition method based on "selective attention" which uses a non-deterministic finite automaton (NFA) and event detection in focusing regions corresponding to its current states. This method, however, has an essential problem in that it doesn't have any object concept and therefore can't recognize multiple objects. We solve this problem by introducing "token" which represents the target object. We, furthermore, propose an extension of this method to improve the ability of recognition by using multi-viewpoint image sequences. In this extension, three types of information-integration schemes are possible at image, event and state levels. Among these schemes, we show that the state-level integration is the most effective one. The effectiveness of the proposed method is demonstrated in experimental results.

1 はじめに

広域分散監視システムにおいて、個々のカメラで得られる映像からシーン中での移動対象の動作を認識することは、最も基本的な問題である。一般に動作認識問題は

1. 運動解析：対象の物理的な運動の解析
2. 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
3. 行為理解：対象の動作からの意図理解

の3つのレベルの問題に大別されるが、本報告では2.の動作識別問題を取り扱う。我々は、特徴抽出と抽出された特徴系列の解析という2つの処理を行う従来のボトムアップ的な動作識別法の問題点を克服するために、トップダウン処理を導入した「選択的注視に基づく動作識別法[1]」を提案している。

しかし、この手法では「イベント系列」の認識を行なっているに過ぎず、動作対象(以下、単に対象)に関する概念が欠落しているため、対象の個数を認識することは本質的にできない。そこでまず、状態遷移モデルにおいて対象を表現するための「トークン」という概念を導入する。これは、1つの対象の動作に対応する複数の遷移状態系列に同一のトークンIDを与え、対象の個数を認識する手法である。

さらに、単一視点の映像のみを用いた場合、画像が3次元対象世界の2次元投影像像であることから、カメラの視線方向に沿った対象の動作が認識できない場合がある。この問題を解決するために、視野を共有する複数台のカメラから得られる多視点映像を用いた協調的な動作認識への拡張が考えられる。その際に、複数の観測ステーションから得られる情報を「映像」「イベント」「状態」のうちどの段階で統合するかによって3種類の統合法が考えられる。このうち「状態レベルでの統合法」が理論的に最も優れていることが示せる。この方法では、各観測ステーションが独立に状態遷移モデルを持つため、異なるステーション間での状態の対応付けが必要になる。この対応付けを行うための情報として、異なる状態遷移モデル間での「状態間の共起関係」が利用できる。共起関係を満足する状態の組が、対象の動作段階に対応するものと考え、動作段階の表現として不適切な 共起性を持たない状態の組を排除する手法を提案する。この場合にもトークン伝搬のアルゴリズムを用いることができ、「対象概念の導入」と「多視点映像の利用」により、安定かつ正確な動作認識が実現できる。

本稿では、まず2章でこれまでの手法を簡単に紹介した後、3章で「対象概念の導入」、4章で「多視点映像への拡張」について論じる。そして、5章では、実際に 本稿の手法を用いて多視点映像による安定な動作認識システムが構成できることを、実験結果に沿って示す。

2 選択的注視に基づく動作識別

一般に、画像を用いた動作識別は、特徴抽出と抽出された特徴系列の解析という2つの処理によって実現することができる。

従来の動作識別法では、特徴抽出部には、直交関数展開係数などの画像のセグメンテーションを行わない手法、特徴系列の解析には、入力系列に対して最も良くマッチするモデルを最適化によって求めるという隠れマルコフモデル(HMM)が一般的に用いられ、これらの処理を順番に実行するボトムアップ処理が行われている。このような構成では、HMM自体がセグメンテーションの機能を持たないため、入力の時間的空間的なセグメンテーションを特徴抽出部で行わなければならない。しかし、一般的にセグメンテーションは不安定であり、その導入によってシステム全体の安定性が損なわれてしまうという問題がある。

このような問題を解決するため、我々は系列の解析結果を特徴抽出にフィードバックするトップダウン処理を導入した、選択的注視に基づく動作識別法を提案した。この手法では、特徴抽出は、ある画像領域(注目領域)内での画像の変化(イベント)検出、系列の解析は、イベント検出結果に応じて状態遷移を起す非決定性有限オートマトン(NFA)によって実現されており、注目領域をNFAの状態に応じて設定することにより、系列の解析結果に依存したトップダウン的な特徴抽出が実現されている。この手法は、動作識別だけでなく、注目領域とNFAの導入による画像の空間的時間的なセグメンテーション機能を有しており、従来法の適用ができない複数動作の同時識別問題に適用することができる。

以下、具体的な機構について述べる。

2.1 動作同定機構

まず、与えられた動画像データが特定の動作クラスに属するか否かを判別する「動作同定機構」について説明する。この機構は、次の3つの要素から構成されている(図1参照)。

注目領域シーケンス: NFAの各状態に対応した動作段階における画像中での固有変化領域を示す。教師あり学習によって作成する。

イベント検出器: 入力された画像中の注目領域における変化を検出し、複数の注目領域についての変化検出結果を結合したイベントコードを出力する。具体的には、背景差分により

求められた変化画素の占有率が$\theta(0 < \theta < 1)$以上のときに変化を検出する。

動作記述用オートマトン: 図1のように順序付けられた状態 $\{q^0, q^1, \cdots, q^m (= q^{acc})\}$ を持つNFA。各状態が対象の動作段階に対応する。

これらの要素間で「現在の状態に対応する注目領域をイベント検出器に入力し、出力されるイベントコードにより、NFA において状態遷移が起きる」という手続きを繰り返す。

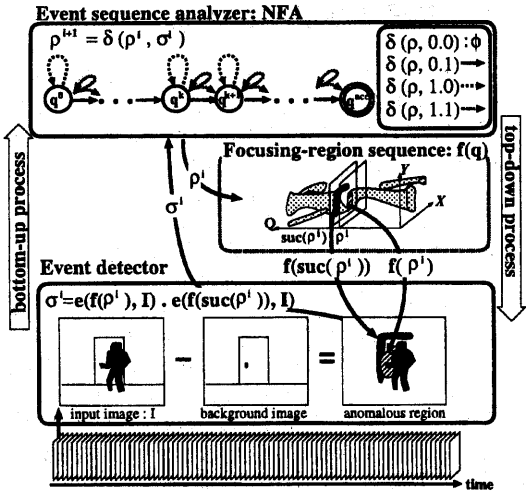

図 1: 動作同定機構

2.2 動作識別機構

複数の動作を識別する動作識別機構は、各動作クラスω_i $(i = 1, \cdots, N)$ に関する動作同定機構に、新たな初期状態 q^0と、q^0から各同定機構の初期状態 $q^0_{\omega_i}$ へのϵ-遷移を付け加えることで構成される (図2)。各同定機構の要素は他クラスのものとは独立に学習可能である。

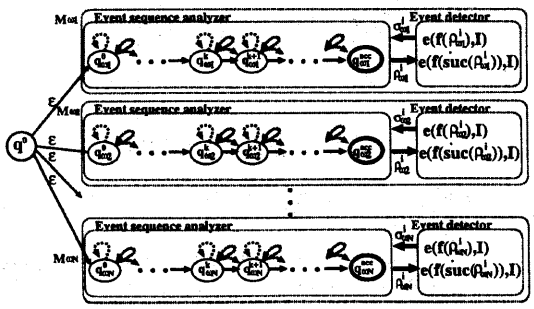

図 2: 動作識別機構

3 対象概念の導入

広域監視システムにおける動作認識においては、複数の対象を分離するため、対象に関する概念を導入することが不可欠である。しかし、前述の動作識別機構の状態遷移モデルにおいては、対象を明示的に取り扱っていないため、対象の個数が本質的に認識できなかった。

そこで、ここでは 状態間の連続性を利用することで対象を表現した「トークン」という概念を導入し、対象個数の認識を可能にする。

3.1 トークン

選択的注視に基づく動作識別においては NFA を用いるために一つの対象の動作段階に対する遷移状態* は複数となる。我々が扱っている対象の動作は「時空間を占める連続的な変化領域」であるため、一つの対象は 時間的に連続した遷移状態の集合に対応する。

したがって、同一の対象に対応することを表現するために、連続した遷移状態に同一 ID を持つ「トークン (token)」を割り当てる。このトークンを遷移状態にしたがって伝搬させ、トークンが最終状態にたどり着くことで、その ID に対応する対象の動作を受理する。

3.2 トークン共有区間

一つの連続遷移状態が必ずしも一つの対象の動作段階に対応するとは限らない†。したがって一般には、一つの遷移状態に対して、異なる ID を持つ複数のトークンを割り当てるべきである。

そこで、連続した遷移状態の集合に対し、一つの「トークン共有区間」C^s $(s = 1, \cdots n)$ を定義する (図3)。共有区間 C^sはその属性として、トークン ID 集合、およびその ID のトークンを共有し得る遷移状態の集合を持つ。このとき、異なる共有区間に同一 ID のトークンは存在しない。すなわち、共有区間 C^sに属するトークン ID の集合を $C^s.id$ とすると

$$(C^{s_1}.id) \cap (C^{s_2}.id) = \phi \quad (s_1 \neq s_2) \qquad (1)$$

これを「トークン分布に関する制約」と呼ぶ。

3.3 トークンの伝搬

以上では、ある時刻での、同一 ID のトークンを持つ遷移状態集合である共有区間について述

*本稿では、現在の状態集合に含まれる状態を「遷移状態」、それ以外を「非遷移状態」と呼ぶ。

†例えば、2 つの対象のうち片方が他方の動作段階を途中で追い抜かす時

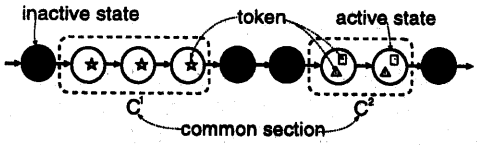

図 3: トークンと共有区間

べたが、ここでは時間の経過に伴うトークンの伝搬法について具体的に述べる。

ある時刻 i において、NFA に共有区間 C_i^s が存在し、各遷移状態に複数のトークンが割り当てられている。そして、時刻 $i+1$ の入力画像から生成したイベントコードにより新たな遷移状態が定まり、その遷移状態の連続性をもとに新たな共有区間 C_{i+1}^t を作成する。

この新たな共有区間 C_{i+1}^t に対して、時刻 i の共有区間 C_i^s が持っていたトークン ID を、「トークン分布に関する制約」を満たすように割り当てる。具体的には、共有区間 C_{i+1}^t に属する状態の遷移元の状態集合を求め、それらが属する旧共有区間 $C_i^{s_1} \cdots C_i^{s_n}$ を調べる。そして、図 4 のように、それらの旧共有区間と新共有区間 C_{i+1}^t との間にリンクを張る。このリンクを通して旧共有区間から新共有区間へとトークン ID を割り当てる。

その際、2 時刻間の共有区間のリンクにより次の 3 つの状況が考えられる。

1. 共有区間が一対一に対応づけられる
2. 時刻 i の複数の共有区間と時刻 $i+1$ の単数の共有区間が対応づけられる (共有区間の融合)
3. 時刻 i の単数の共有区間と時刻 $i+1$ の複数の共有区間が対応づけられる (共有区間の分裂)

1. の場合は、旧共有区間のトークンをそのまま受け継ぐ。

2. の場合、時刻 i では離れた動作段階にあった複数の対象が、時刻 $i+1$ で近い動作段階に移ったと考えられる。この場合、融合前の複数の共有区間に属するトークン ID を融合後の共有区間がすべて受け継ぐ。

3. の場合は 2. の逆であるから、分裂前の共有区間の持つ n_s 個のトークン ID を、分裂後の d_s 個の共有区間に適当に分配する。ただし $n_s < d_s$ であれば、不足分のトークン ID を新たに生成する。

以上の操作により、C_{i+1} のトークン ID が定まる。これらの ID を持つトークンを、C_{i+1} に属する状態集合内に伝搬させ、時刻 $i+1$ におけるトークンの伝搬が完了する。

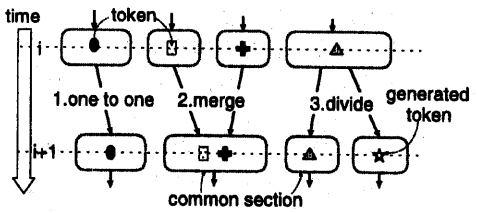

図 4: リンクを介したトークン ID の割り当て

4 多視点映像への拡張

これまで述べてきた単一視点からの映像を用いた認識機構では、動作の種類によっては認識能力が低下する場合がある。これにはカメラの視線方向と対象動作の方向とのなす角度が関係する。この角度が小さい場合、動作が進行しても画像中における変化領域がそれほど大きく変動しないため、動作段階の特定・複数対象の分離 が困難になり、結果として認識能力が低下する。

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) を補いあ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正確に把握することができる (図 5)。

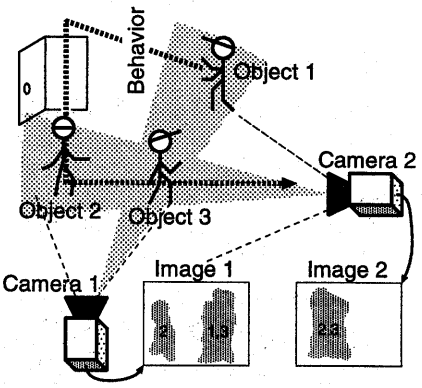

図 5: 複数カメラの使用

複数カメラを使用するにあたって、次の 2 つの前提を置く。

- シーンやカメラ間の空間的位置・方位の情報は得られていない (Uncalibrated)
- 各カメラ間での画像の撮影時刻は合致している

前者は実際の監視システムにおいて一般的な状

況であり、ここでの複数カメラの利用が、3次元世界を復元するステレオ視とは異なることを意味する。また、後者は、複数の映像による動作認識を行なうための重要な条件である。

複数のカメラを用いる場合、複数の情報(映像)が得られるが、それを協調させて一つの結果を出すためには、前述の認識機構のいずれかの段階で情報を統合する必要がある。その段階には

1. 映像レベル
2. イベントレベル
3. 状態レベル

の3つが考えられ、各々に対して統合法が考えられる(図6)。以下では、1,2.の統合法について簡単に述べた後3.の統合法について論じ、多視点映像を用いる意味を明らかにする。

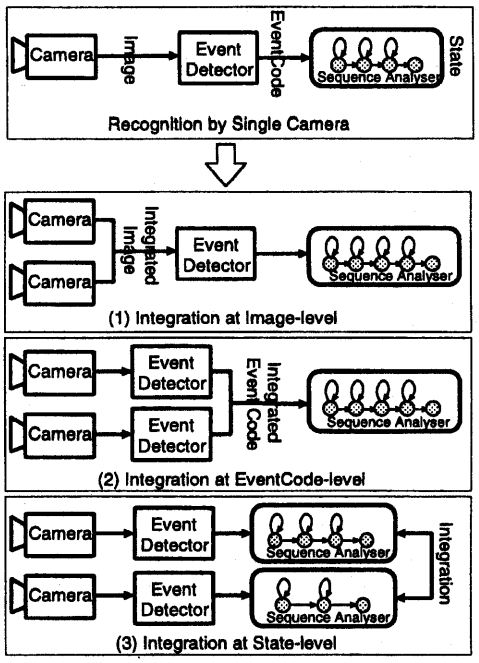

図 6: 3種類の統合法

4.1 映像レベルでの統合法

まず最も単純な方法として、複数のカメラで得られた画像 $I_c(c=1,\cdots,N_c)$ の張り合わせ画像 I_{all} を一つのカメラで得られた一枚の画像とみなして、従来の単一映像による認識法をそのまま用いることが考えられる。これを映像レベルでの統合法と呼ぶ。

しかし、この方法では、異なるカメラパラメータによって得られた画像 I_c を、すべて一様な入力として扱っているため、違なった意味付けの量を加えあわせているという問題がある。

4.2 イベントレベルでの統合法

次の方法は、観測ステーションが独立にイベント検出器を持ち、出力されるイベントコード e_c をもとに一つのイベントコード e_{all} を生成して、それを共通のオートマトンに入力する方法である。これをイベントレベルでの統合法と呼ぶ。

e_{all} の具体的な求め方としては、イベント検出結果の論理積を用いる(すなわち $e_{all}=\bigcap e_c$ とする)方法が考えられる。論理積を用いることは、各画像での注目領域の3次元空間中での重なりを扱う効果が生まれ、認識能力の向上が期待できる。

しかし、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」という非常に強い条件を課すことになり、時空間における対象動作の僅かな変動も許容しない認識機構になる。N_c が増えるに従ってこの傾向は強まる。

4.3 状態レベルでの統合法

最後の方法は、観測ステーションが各々独立に状態遷移モデル NFA_c を持ち、それらの内部状態の情報を統合することによって一つの結果を導くものである。これを状態レベルでの統合法と呼ぶ。

前述のイベントレベルでの統合法では、1つのNFAによって動作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。これは、状態レベルでの統合法において、全く同じ構造のNFAを用いる場合と等価である。このことから、イベントレベルでの統合法が、状態レベルで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。

ここでは、状態レベルでの統合法に必要な状態の対応付けを行なうための「状態の直積空間」について説明し、その空間内の「許容経路」と呼ばれる一本の経路に沿ったトークンの伝搬を考えることにより、動作認識が可能となることを述べる。その際に、状態間の共起性を用いることで、複数対象を分離する能力が向上し、対象の動作段階に関する曖昧性を低減できることを示す。さらに、許容経路を拡張することによって、対象の動作変動を許容する手法を提案する。

4.3.1 状態の直積空間

各NFAの個々の動作同定機構の状態は、順序付けられている。このことを利用すると、個々の

動作ごとに、各NFAの状態を基底とした「状態の離散直積空間 (以下、単に直積空間)」を張ることができる。ここで、各NFAの状態間には、特定の動作段階に対して同時に遷移状態になるべき「共起性」を持つ組が存在する。直積空間中での点として表わされる この組を「許容状態組」と呼ぶ。学習時に得られる各状態の絶対時間の情報を利用すると、許容状態組の順序集合が得られ、その要素は直積空間中で一本の経路を形成する。これを「許容経路」と呼ぶ (図7)。

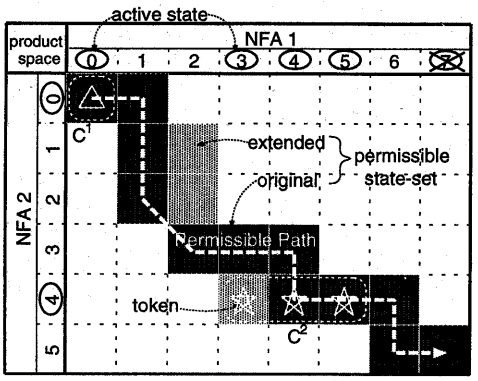

図 7: 状態の直積空間 ($N_c = 2$)

許容経路は、図8のような N_c部グラフとしても表現可能である。この N_c部グラフでの順序付けられたクリークが、直積空間内での許容状態組に対応する。

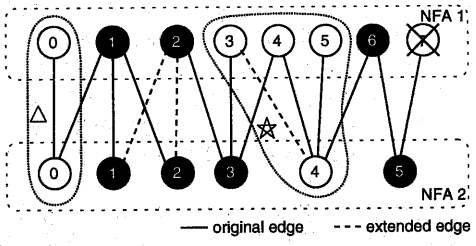

図 8: 直積空間に対応する $N_c(2)$ 部グラフ

直積空間内では、順序付けられた各許容状態組が、対象の動作段階を表現する。これは、これまで考えてきた「状態 – 動作段階」の対応を、「状態組 – 動作段階」へと拡張することを意味するが、これまで単独のNFA上で考えていたトークンの伝搬を許容経路上の概念に拡張することにより、動作認識を行なうことが可能である。つまり、共起性を持つ遷移状態の組 (許容遷移状態組) の許容経路上での連続性をもとに、許容経路上でトークンを伝搬させる。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次元中でのより限定した領域に注目することができるため、複数対象の分離能力が向上する。

ここで、各NFAの遷移状態のうち許容遷移状態組に関与しないものは、動作段階を表現するのに不適切なものとして遷移状態から削除する。これは、状態間の共起性を用いたフィルタリングであり、単一視点映像による動作認識で現れる 動作段階の曖昧性の低減に効果がある。

4.3.2 動作変動の許容

直積空間において、絶対時間の情報を用いて許容経路を作成した場合、理想的なデータが入力されたときは正しく動作が認識されるが、これではイベントレベルでの統合法において論理積を用いた場合と同様で、入力データの変動への対応ができない。

そこで、入力データの変動による遷移の失敗を防ぐため、許容経路付近の 絶対時間を共有しない状態組を 新たに許容状態組に追加し、許容経路を拡張することを考える (図7参照)。これは、動作が「時空間中での連続的な変化領域」であることを考えれば、「時空間中でのわずかな変動」を時間方向において許容することで、動作変動への適応性を高めることを意味する。

許容経路の具体的な作成法としては、次の2つが考えられる。

- 絶対時間の情報を利用せずに、サンプルデータを入力した際の共起性を直積空間に投票し、その結果をもとに、直接 許容経路を求める
- 絶対時間の情報を利用して求めた許容経路をもとにして、適当な方法を用いて許容経路の拡張を行なう

しかし、前者の方法では、入力に対しNFAに複数の状態遷移が起きる (すなわち複数の内部時間が流れる) ため、許容経路の作成を安定に行なうことが困難であるので、後者の方法が適当である。具体的には「サンプルデータを入力した際の許容経路上のトークンの伝搬を追跡し、伝搬が途切れる点付近で、各NFAの遷移状態に基づいて許容状態組を追加する」という操作を行なう。

本節で述べた「状態レベルでの統合法」は、各ステーションの協調という観点から見れば、最も分散度の高い協調法である。また、特殊なケースとして他の2つの統合法を表現でき、状態空間中での許容経路の決定に対して自由度があるという点で、今後の拡張性も大きい。

5 実験

ここでは、本稿で提案した動作認識システムを実際に構成し、動作認識を行なった結果を示す。

実験環境としては、部屋に2つのカメラ(1,2)を設置し、定めた経路に沿って人が部屋に出入りするシーンを撮影して取り扱う(白黒256階調 320 × 240、30[フレーム/秒])。動作クラスは「入室」「退室」の2クラスである。学習用サンプルデータとしては各々の動作クラスに対し、20個の映像を撮影する。これを背景差分して求めた変化領域から注目領域を学習する。入室動作の原画像・変化領域・注目領域の一例を図9に示す。また、認識対象の画像の一例を図10に示す。

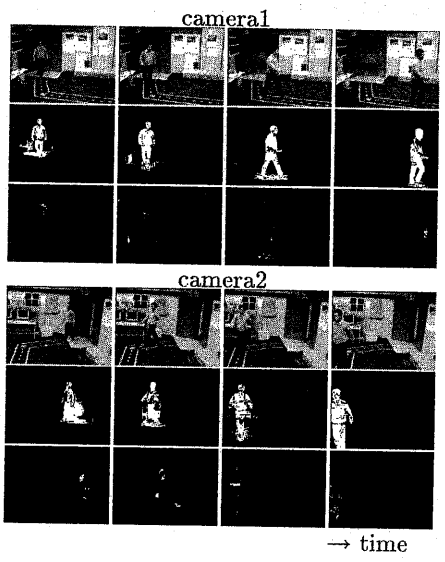

図 9: 入室動作 (原画像、変化領域、注目領域)

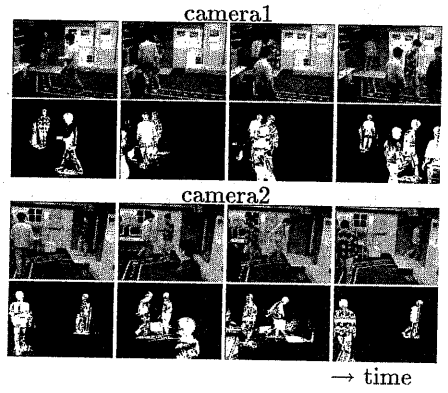

図 10: 認識対象画像の一例 (原画像、変化領域)

ここでは2台のカメラが利用できるので、次の5種類の動作認識機構が構成できる。

S1: カメラ1のみを使用

S2: カメラ2のみを使用

DI: (カメラ2台を用いて) 映像レベルの統合法

DE: イベントレベルの統合法

DS: 状態レベルの統合法

このうち、DSを用いてあるシーン(2人が「入室」する)を認識したときの各NFAでの状態遷移の様子を図11に示す‡。単独のNFAでは、灰色の部分が動作段階に関する曖昧性として現れていたが、方法DSではこの部分の状態を削除することにより曖昧性が低減でき、対象個数の認識に有効であることが分かる。

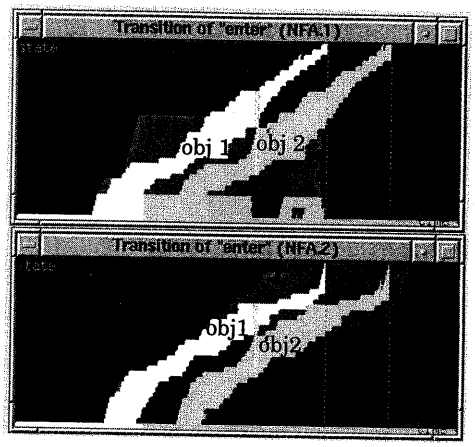

図 11: 状態遷移の例 (縦軸:状態、横軸:時間)

各方法の認識能力について調べる。我々は広域監視での動作認識に特有な「複数の対象」の存在するシーンでの認識能力の向上を目指している。そこで、もっとも単純な場合として、2つの対象が同時に動作を行なう次の3種類のシーンを、各々20個ずつ撮影する。

1. 二人が「入室」を行なう
2. 二人が「退室」を行なう
3. 一人が「入室」、一人が「退室」を行なう

各認識機構にこれら計60個のデータを認識させたときの正解数を、イベント検出器での閾値θを変化させて調べた。結果を図12に示す。

個々の認識法においては、θが大きすぎると状態遷移が途中で途切れやすくなるため、θが小さ

‡ここでは、S1,S2との比較を行なうために、共起性が認められない遷移状態の削除を行なっておらず、その部分を灰色で示す。実際にはDSでは遷移状態の削除を行なう。

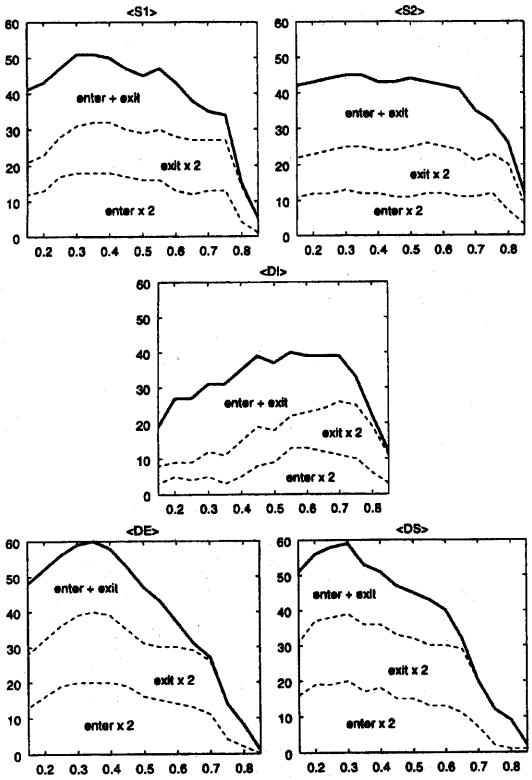

図 12: 認識結果 (縦軸:正解数、横軸:閾値θ)

すぎると複数対象を分離しにくくなるために、認識率が悪くなる傾向が見られる。方法によって最適なθは異なるが、各方法での最良値を比較すると、2 人が同一動作を行なうシーン (1,2) に対する認識能力に関しては、DI では S1,S2 よりもかえって悪化するが、DE,DS では大幅に改善され、多視点映像の利用の有効性が認められる。ただし、2 人が異なる動作を行なうシーン (3) に関しては、特に認識能力の向上は見られない。つまり、多視点映像への拡張は、同一動作を行なう対象の個数の認識に対して、大きな効果が得られるといえる。

ただ、今のところ、理論的に最も優れている DS の認識能力が DE よりも僅かに低い。これは

- 直積空間での許容経路の拡張法がまだ不完全
- カメラ数が多くないため、DE でも動作変動への適応性が低くならない

などの理由によると考えられる。

6 おわりに

本稿ではまず、広域監視システムにおける動作認識法である文献 [1] の「選択的注視に基づく動作識別」を紹介した。しかし、この機構では 視覚監視の動作認識に不可欠な 対象に関する概念が欠落していた。そこで 我々は、対象を表現する「トークン」という概念を導入し、対象個数の認識を可能にした。次に、動作クラスに依存した認識能力の低下を解決するため、多視点映像を用いた協調的動作認識法への拡張について論じた。その際に、情報を統合する段階によって 3 つの方法が考えられ、そのうち 分散度の高い「状態レベルでの統合法」が最も本質的・一般的で今後の拡張性も高いことを述べた。そして、実験結果から多視点映像への拡張の有効性が示せた。

今後の課題としては以下の点が挙げられる。

1. 動作変動の空間方向での許容
2. 分岐構造を持つ NFA への拡張
3. 対象動作との相関が小さい変化 (ドアの開閉など) が起こる場合の注目領域の学習法
4. 複数カメラで絶対時間が得られない場合の状態間の対応付け
5. 光環境の変化などに対してロバストな変化検出オペレータ
6. 首振りカメラへの拡張・状況に応じたカメラアクション
7. ステーション間でのメッセージ通信

特に 1,2. に関しては、本稿で提案した手法の動作変動への適応性には限度があり、それを克服するという意味で非常に重要である。また、3. に関しては、特に屋内の監視において一般的な状況であるので、汎用性のある視覚監視を行なうためには不可欠である。今後これらの点を一つずつ解決することにより、信頼性のある広域監視システムの確立を目指す。

参考文献

[1] 和田, 加藤: 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, CVIM 103-6 (1997)

[2] J. Hopcroft, J. Ullman 著 (訳:高橋正子 他), “オートマトン言語理論計算論 1”, サイエンス社 (1979)