

PrefixSpan を用いた人物の日常行動抽出

戸成 拓也[†] 滝口 哲也^{††} 有木 康雄^{††}[†] 神戸大学大学院工学研究科 兵庫県神戸市灘区六甲台町 1-1^{††} 神戸大学自然科学系先端融合研究環 兵庫県神戸市灘区六甲台町 1-1E-mail: [†]tonaru@me.cs.scitec.kobe-u.ac.jp, ^{††}{takigu,ariki}@kobe-u.ac.jp

あらまし 従来の動作認識では、歩く、走る、手を振るなど単純な動作に対して研究が行われている。しかし、映像から人物の動作に基づいて日常行動を解析するには、複雑な動作を認識する必要がある。そこで本研究では、人物の行動を単位動作記号列として表し、出現頻度の高い記号列を行動として抽出する手法を提案する。まず、映像における人物の動作を単位動作記号列に変換する。これには、教師なし学習による動作認識を用いることで自動的に行うことができる。次にシーケンシャルパターンのマイニング手法である PrefixSpan を用いて、単位動作記号列から頻度の高い記号列を抽出することで、日常行動を抽出する。実験には、研究室で撮影したデスクワーク映像 70 分を使用した。実験より、人物の行動を抽出した結果、平均再現率 86.0%、平均適合率 78.3% であった。

キーワード 人物行動, 動作認識, pLSA, prefixspan

Extraction of Human Daily Activities from videos as Action Sequences using PrefixSpan

Takuya TONARU[†], Tetsuya TAKIGUCHI^{††}, and Yasuo ARIKI^{††}[†] Graduate School of Engineering, Kobe University 1-1, Rokkodai, Nada, Kobe, Hyogo^{††} Organization of Advanced Science and Technology, Kobe University 1-1, Rokkodai, Nada, Kobe, HyogoE-mail: [†]tonaru@me.cs.scitec.kobe-u.ac.jp, ^{††}{takigu,ariki}@kobe-u.ac.jp

Abstract Conventional action recognition was dedicated to the single action such as walking, running and waving hands. However, recognition of human activities based on human actions requires the complicated analysis. From the viewpoint, we propose a framework for recognizing human activities in our daily life after human action recognition. Since a human activity is represented as a sequence of actions, the actions are recognized from videos and then frequently-occurring human activities are extracted from them. At first, human actions are recognized by unsupervised action learning and then human activities are extracted from a sequence of actions using PrefixSpan, which is a kind of mining techniques of sequential patterns. We show the experimental results applied to the data taken in a deskwork environment to demonstrate the performance of the proposed framework. As the experimental result, 86.0% averaged recall rate and 78.3% averaged precision rate were obtained in extracting the human activities.

Key words Human activity, action recognition, pLSA, prefixspan

1. はじめに

近年、主記憶装置の大容量化などで個人の日常行動を映像として残すことが容易になってきている。この日常行動の映像から人物の行動を解析することで、ユーザーの生活を改善するための情報の提示などが期待できる。例えば、デスクワークをしている人物がいるとする。彼は PC で作業をしているが、しばしばコーヒーを飲んだり、音楽を聴くためにヘッドフォンをつけたりしている。もし彼がコーヒーを飲みすぎていると、シス

テムは彼に健康上での警告を発することができるだろう。しかし、行動を解析するには、映像中で人物がいつどのような行動をとったかを理解する必要がある。そのためにはまず、映像中の人物の行動を認識する必要がある。また事前にラベル付き学習データを用意することは実用上望ましくないため、行動の認識は自動で行なえるほうが望ましい。

現在、走る、歩く、手を振るといった単純な動作認識の研究が広く行われている [3] [4] [5]。例えば、Laptev ら [3] は映像の時空間上に現れる Interest Point を特徴量とし、SVM を用い

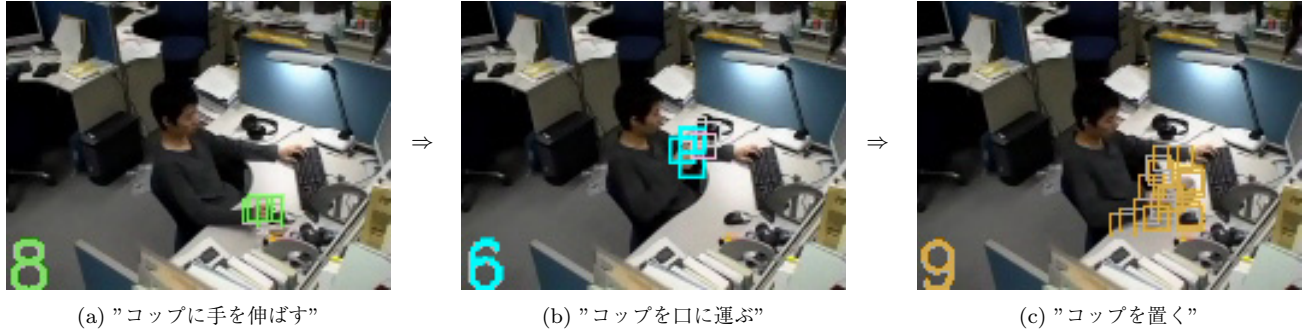


図1 “コーヒーを飲む”という行動を構成する単位動作。左下の数字は単位動作記号を表す。

て動作認識を行っている。また Niebles ら [5] は pLSA モデルを用いて、教師なし学習と複数の人物の動作認識を提案している。しかしながら、人物の行動は様々な動作の組み合わせで成り立っており、これらの手法をそのまま適用するだけでは、複雑な動作を認識することはできない。

また関連する研究として、人物行動を階層構造を持つ行動の記号列として表し、行動を解析する研究がある。代表的なアプローチでは、Stochastic Context-Free Grammar(SCFG)を用いて人物の行動を解析している [7] [8] [9]。しかし、これらの研究では文法は事前に与えられている。これに対し木谷ら [11] は、最小記述長原理に基づきノイズを除外した終端記号集合と、それに伴う文法の獲得手法を提案している。また、Hamid ら [10] はキーオブジェクトとのインタラクションの記号列が、SuffixTree で表せることを利用して、台所における人物の行動の認識を行っている。これらの研究では、1つの単位行動やキーオブジェクトとのインタラクションを記号に置き換えることで、人物の行動解析に自然言語処理の研究を持ち込むことに成功している。

我々の目標は、自動で映像中の人物の行動を、同じものをまとめて抽出することである。そうすれば、ユーザーは得られた行動の集合にラベルを付けるだけになり、事前の学習データの収集に関するユーザーの負担を減らすことができる。また、従来では行動を構成する単位要素は別処理で行われていたが、これも映像から自動的に得ることが望ましい。そこで我々は、映像中の人物の行動を抽出するために、行動を単純な動作の記号列として扱うことで行動を抽出する手法を提案する。さらに、行動を構成する単純な動作の記号列の取得に関しても、自動で行うことを試みる。人物の日常行動は、頻度が高く発生すると仮定することで、行動を表す記号列が、映像から得られた動作記号列に多く出現すると考えられる。提案する手法は2つの段階で構成されている。まず、Human Action Categorizing Method [5] を用いて、映像を自動で単位動作の発生確率シーケンスに変換する。次に、得られた単位動作の発生確率シーケンスを、離散化された単位動作記号列に変換し、PrefixSpan を用いて頻度の高い記号列を行動として抽出する。本稿は5章から構成される。第2章では本研究での行動を定義する。第3章では提案手法を述べる。第4章ではその有効性を検証し、第5章で問題点、将来の課題についてまとめる。

2. 提案手法

本章では提案手法について述べる。提案する手法は2つの段階で構成されている。1つ目は、Human Action Categorizing Method [5] を用いて、映像を単位動作の発生確率シーケンスに変換する。これは、bag-of-words な時空間特徴量と pLSA を用いることで、教師なし学習を行う。2つ目は、得られた単位動作の発生確率シーケンスを、離散化された単位動作記号列に変換し、PrefixSpan を用いて頻度の高い記号列を行動として抽出する。PrefixSpan はシーケンシャルパターンマイニングの代表的な手法であり、効率的に頻出する部分記号列を抽出することができる。

2.1 行動の定義

本節では本研究での行動を定義する。人物の行動は様々な動作から成り立っており、それぞれを1つの記号で表現すれば、行動を動作の記号列で表すことができる。例えば、以下のようなコーヒーを飲むという行動の記号列 S が与えられたとする。

$$S \longrightarrow 8 \ 6 \ 9$$

各単位動作 8, 6, 9 はそれぞれ、“コップに手を伸ばす”、“コップを取る”、“コップを置く”を表す。その様子を図1に示す。コーヒーを飲むという行動はコップに手を伸ばし、コップを口に運び、そして飲み終わったらコップを置くという、一連の流れで成り立っている。つまり、行動の記号列は単位動作記号を時系列で並べたものである。

2.2 Human Action Categorizing Method

Human Action Categorizing Method は時空間特徴を抽出し、pLSA モデルを用いて教師なしで学習・認識を行う手法である。本節では、この手法の詳細について述べる。

2.2.1 特徴量抽出

映像の画素毎に、輝度変化の激しさを返すレスポンス関数を、以下のように与える。

$$R(x, y) = \left(I(x, y) * g(x, y; \sigma) * h_{ev}(t; \tau, \omega) \right)^2 + \left(I(x, y) * g(x, y; \sigma) * h_{od}(t; \tau, \omega) \right)^2 \quad (1)$$

ここで $I(x, y)$ は、映像中のグレースケールのフレーム画像を表す。 $g(x, y; \sigma)$ は $I(x, y)$ を平滑化するための2次元ガウシアンフィルタである。また、 $h_{ev}(t; \tau, \omega)$ と $h_{od}(t; \tau, \omega)$ は1次元

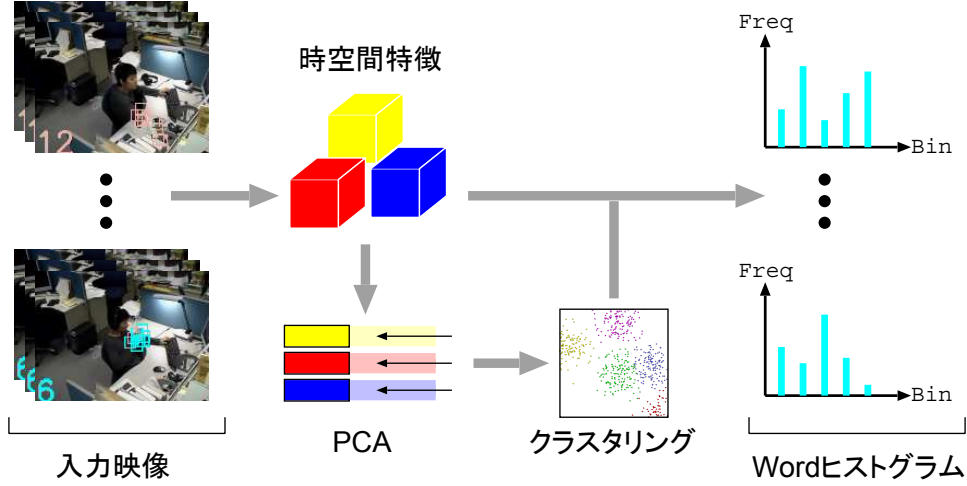


図 2 特徴量抽出の流れ

Gabor フィルターであり、それぞれ以下のように与えられる。

$$\begin{aligned} h_{ev}(t; \tau, \omega) &= -\cos(2\pi t\omega)e^{-\frac{t^2}{\tau^2}} \\ h_{od}(t; \tau, \omega) &= -\sin(2\pi t\omega)e^{-\frac{t^2}{\tau^2}} \end{aligned} \quad (2)$$

1次元 Gabor フィルターは、映像の時間軸 t に沿って畳み込まれる。2つのパラメータ σ と τ は、それぞれのフィルターのスケールを表す。また ω は、[5] において $\omega = 4/\tau$ を効果的な値としている。

レスポンス関数は、映像上の複雑な動きが発生した領域を検出する。複雑な動きを含む領域に対して大きな値を返し、あまり複雑ではない動き、例えば単純な平行移動運動に対しては、あまり大きな値を返さない。次に、レスポンス値が空間上で極大値をとる点を中心に、映像から画素をキューブ状に切り出す。切り出すサイズは、[5] においてフィルターのスケールパラメータ σ, τ の約 6 倍程度が推奨されている。切り出されたキューブは、画素毎に輝度勾配に変換し、さらに 1 本のベクトルに変換する。そして、PCA を用いて各特徴ベクトルを次元圧縮する。

得られた特徴ベクトルを k-means でクラスタリングすることにより、それぞれ各クラスターに分類される。それゆえ、映像中で発生した特徴ベクトルの集合は、各クラスターをピンとするヒストグラムで表される。以降、クラスタータイプを word と呼ぶ。特徴量抽出の流れを図 2 に示す。

2.2.2 pLSA による単位動作の学習・認識

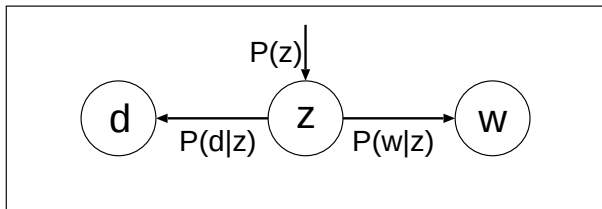


図 3 対称にパラメータ化された pLSA モデルのグラフィカルモデル

pLSA (Probabilistic Latent Semantic Analysis) は、共起データの解析に用いられている手法である。ここでは、映像で

発生した word の共起を基に、単位動作に対応する隠れトピックの混合によって映像を表現する。

2.2.1 で述べた特徴抽出を用いて、人物の動作を含む映像 $d_j \in D = \{d_1, \dots, d_N\}$ と word $w_i \in W = \{w_1, \dots, w_M\}$ の間の共起テーブル $\bar{N}$ を用意する。さらに、隠れトピック変数 $z_k \in Z = \{z_1, \dots, z_K\}$ を用意する。隠れトピックは各単位動作に対応するため、隠れトピックを単位動作と呼ぶ。ここで、単位動作 z_k が発生したとして、観測される (w_i, d_j) のペアがそれぞれ独立に生成されるとすると、その同時確率は、

$$P(w_i, d_j) = \sum_{k=1}^K P(z_k)P(w_i|z_k)P(d_j|z_k) \quad (3)$$

のように表現される。 $P(w_i|z_k)$ は単位動作 z_k における word w_i の発生確率であり、 $P(d_j|z_k)$ は単位動作 z_k における映像 d_j の発生確率である。このモデルは対称にパラメータ化された生成モデルであり [1]、そのグラフィカルモデルを図 3 に示す。

式 (3) の確率分布を得るために、各パラメータ $P(z)$, $P(w|z)$, $P(d|z)$ を式 (4) の対数尤度が最大となる基準で求める。

$$L = \sum_{i=1}^M \sum_{j=1}^N n(w_i, d_j) \log P(w_i, d_j) \quad (4)$$

ここで、共起テーブル $\bar{N}$ の要素 $n(w_i, d_j)$ は、映像 d_j における word w_i の発生個数である。対数尤度最大化を解くために、Expectation Maximization (EM) アルゴリズムを用いる。

次に、得られた確率分布を用いて認識を行う方法について述べる。テスト映像 d_{test} が与えられたとすると、その単位動作 z_k の事後確率は以下のように与えられる。

$$P(z_k|w_i, d_{test}) = \frac{P(w_i|z_k)P(z_k|d_{test})}{\sum_{l=1}^K P(w_i|z_l)P(z_l|d_{test})} \quad (5)$$

しかし、 $P(z_k|d_{test})$ は観測されていない。よって、 $P(z_k|d_{test})$ を求めるために、学習フェーズと同じように EM アルゴリズムを用いてこれを求める。ここでは、EM アルゴリズムの計算コストを減らすために逐次型近似 EM アルゴリズムを用いる [2]。

$$P(z_k|d_{test}) = P(z_k|h_N) \quad (6)$$

$$P(z_k|h_n) = \frac{1}{n+1} \frac{P(w_n|z_k)P(z_k|h_{n-1})}{\sum_{l=1}^K P(w_n|z_l)P(z_l|h_{n-1})} + \frac{n}{n+1} P(z_k|h_{n-1}) \quad (7)$$

$$P(z_k|h_1) = P(z_k) \quad (8)$$

$h_n \in H = \{h_1, \dots, h_N\}$ はテスト映像で発生した N 個の word である。

2.3 行動の抽出

本節では、上述の手法を用いて映像を単位動作記号列に自動的に変換し、PrefixSpan を用いて頻度の高い記号列を抽出することで日常行動を抽出する方法について述べる。

2.3.1 単位動作の確率シーケンスへの変換

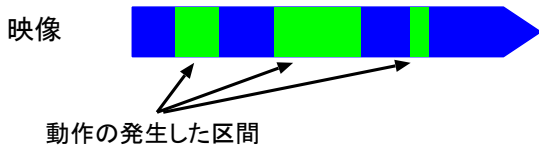


図4 映像から動作が発生した部分映像の切り出し

まず、図4に示すように、動作が発生した区間を映像から切り出し、複数の部分映像を用意する。映像で動作が発生したかどうかを判別するには、式(1)によって得られたレスポンス値の最大値が閾値を超えたかどうかで判断することができる。しかしながら、映像中で動作が発生した部分のみを、そのまま切り出してしまうと、単位動作が連続している場合には、複数の動作をまとめて切り出してしてしまう。しかし、これは問題とはならない。なぜなら pLSA はマルチトピックを扱うモデルであるため、部分映像に複数の単位動作が含まれていたとしても、その映像には複数の単位動作が含まれていると学習されるからである。pLSA による学習後、映像をフレーム毎に d_{test} と置き、そのフレームで発生した word を $h_n \in H = \{h_1, \dots, h_N\}$ として $P(z_k|d_{test})$ を求めることで、フレーム毎の単位動作発生確率が得られる。つまり、映像から単位動作の発生確率シーケンスが得られたことになる。

pLSA で学習を行う際、認識する単位動作数 K を決定する必要がある。もし単位動作数 K が大きい場合、記号化の語彙数は多くなるが、発生する word の少し変化で、異なる単位動作と認識されるため、ノイズなどの影響で認識が安定しにくくなる。逆に単位動作数 K が小さい場合、ノイズに強くなるが、記号化の語彙数は少なくなってしまう。単位動作数 K を自動で設定する方法は、今後の検討課題である。

2.3.2 単位動作の確率シーケンスから単位動作記号列の集合への変換

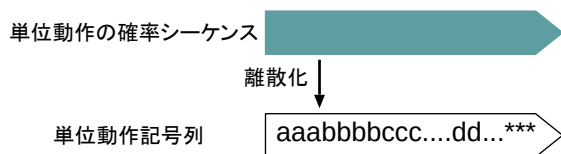


図5 確率シーケンスから単位動作記号列への変換

映像から得られた単位動作の発生確率シーケンスを、離散化された記号列に変換する。まず、ノイズ除去のために確率シーケンスを平滑化する。そしてフレーム毎に確率が最大の単位動作を記号に置き換える。その様子を図5に示す。

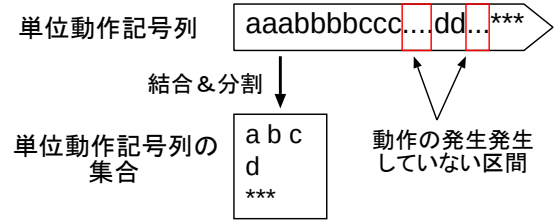


図6 単位動作記号列から単位動作記号列の集合への変換

次に、連続する同じ記号は1つに結合する。さらに動作のない区間が設定した閾値より長い場合、そこで記号列を分割する。その様子を図6に示す。

2.3.3 単位動作記号列の集合から頻出する部分記号列の抽出

人物の日常行動は、高い頻度で発生すると仮定できる。そこで単位動作記号列から頻度の高い記号列を抽出するために、シーケンシャルパターンマイニングと呼ばれる手法を適用する。シーケンシャルパターンマイニングとは、与えられた記号列の集合から、最小サポート値と呼ばれる閾値以上の頻度で現れる部分記号列を求める手法である。シーケンシャルパターンマイニングの手法はいくつかあるが、本研究では効率的な手法の PrefixSpan [6] を適用して、出現頻度順に記号列を抽出する。PrefixSpan は、出現する部分記号列に射影操作を再帰的にを行い、出現する部分記号列とその頻度を得る手法である。射影とは、特定の記号列に続く接尾辞を取り出す操作である。その例を図7に示す。まず4つの記号集合から、単位動作記号の頻度を求め、最小サポート値以上のものを取り出す。次に、その単位動作記号に続く接尾語を取り出し、その集合から最小サポート値以上の単位動作記号を取り出す。この処理を繰り返すことで、出現する部分記号列とその頻度を得る。

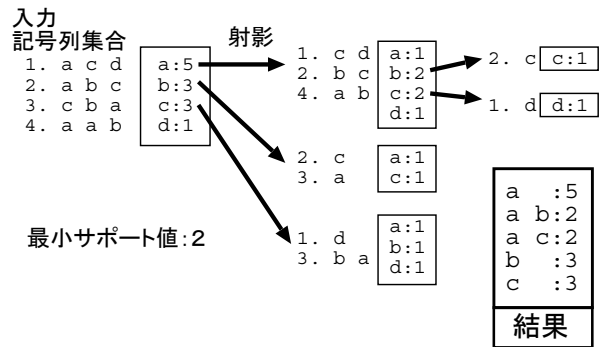


図7 PrefixSpan による頻出する部分文字列の抽出

映像から得られた単位動作記号列集合から、PrefixSpan を適用して頻出する部分記号列を抽出する。PrefixSpan は、単一の記号や、部分記号列関係にある記号列も同時に抽出するため、出現する部分記号列が多く出力されてしまう。そこで、単一の

表 1 PrefixSpan によって得られた部分記号列

頻度	記号列		
16	6 9	5	10 3 7
13	9 11	4	8 11 6 11 9 11 2
12	8 6	4	7 5
11	10 3	4	5 8
10	9 11 2	4	4 13
10	4 10	4	4 10 3 7
10	11 6	4	3 12
9	8 6 9	4	2 5
9	8 11 6	4	13 4
7	6 11 9	4	11 6 9 11
7	4 10 3	4	1 5
6	9 2	3	9 5
6	6 9 11	3	8 6 9 2
6	6 11 9 11	3	8 11 6 9 11 2
6	3 7	3	6 12
5	8 12	3	5 11
5	8 11 6 11 9 11	3	4 7
5	6 9 2	3	4 10 7
5	6 9 11 2	3	13 10
5	6 11 9 11 2	3	12 10 9
5	5 1	3	11 12
5	12 13		

記号や、同じ頻度で部分記号列関係にある記号列は短い方を削除する。次に、得られた頻出部分記号列が行動を表すかどうかをユーザーに提示して、行動を表していればラベル付ける。

3. 実験

本章では提案手法の有効性を検証するために、研究室で 1 人の人物がデスクワークをしている状態を撮影した 70 分の映像を用いる。映像では、人物が PC で作業をしながら、コーヒーを飲む、ヘッドフォンをつける／はずす、ティッシュをとる／捨てる、などの行動を行っている。映像中では他の人物は写らず、また本人も席を離れることはない。映像の解像度は 160 x 120 である。

3.1 実験条件

時空間特徴を抽出するときのスケールパラメータを、それぞれ $\sigma = 11$, $\tau = 19$ とし、クラスタリングにおけるクラスター数 (word 数) を 400 とした。単位動作数 K を 13 とし、PrefixSpan における最小サポート値を 3 とした。また、動作の発生していない区間において記号列分割を行うときの閾値は 120 フレームとした。

3.2 実験結果

PrefixSpan によって抽出された記号列は 43 種類で、そのうち 6 つが行動を表していた。それぞれを表 1, 表 2 に示す。行動を表す部分記号列に対して、それ以外の湧き出しが多く出力される結果となった。行動を表していない部分記号列は、主に行動を表す部分記号列の一部分であった。特に、核となる行動に続いて発生する単位動作、例えばコーヒーを飲み終わったあと、手をコップからマウスに移す動作は、続けて起こることが多く、それを含む部分記号列が多く湧き出した。また、表 2 に

表 2 行動を表している部分記号列

行動	頻度	記号列	再現率	適合率
コーヒーを飲む	16	6 9	1.00	0.91
	7	6 11 9		
ヘッドフォンを外す	7	4 10 3	0.86	0.86
ティッシュを取る	5	8 12	0.80	0.80
頭を掻く	4	4 13	0.50	0.67
ヘッドフォンを付ける	3	4 7	1.00	0.86
	3	4 10 7		
ティッシュを捨てる	3	12 10 9	1.00	0.60

おいて、同じ行動が異なる 2 つのシーケンスとして現れた。これは、図 8(a)(b) の真ん中の単位動作が腕の速度が遅いときに発生するため、同じ行動であっても必ずしも検出されるとは限らないためと考えられる。実験結果より、平均再現率 86.0%、平均適合率 78.3% であった。再現率、および適合率の定義は以下の通りである。

$$\text{再現率} = \frac{\text{True positive}}{\text{True positive} + \text{False negative}} (\times 100\%) \quad (9)$$

$$\text{適合率} = \frac{\text{True positive}}{\text{True positive} + \text{False positive}} (\times 100\%) \quad (10)$$

また、true positive, false positive, false negative の定義を以下のように与える。

True positive: 正しく抽出された行動の数

False positive: 誤って抽出された行動の数

False negative: 抽出されなかった行動の数

Hamid ら [10] は、SuffixTree を用いて台所における行動のクラス発見およびクラスタリングを行い、実験の結果、台所での 7 つの行動を発見し、平均再現率 82.1% 平均適合率 96.2% を得ている。対象とする台所での行動は、夕食を作る、調理後のキッチン片付けなど、複雑であるが記号列は長く、与えられたデータセットもあらかじめ行動ごとの記号列に分割されている。前提や条件は異なるため単純な比較はできないが、今回の結果は、適合率が少し低いものの、良好な結果を示していると考えられる。

4. まとめ

本稿では人物の行動を単位動作記号列として表し、出現頻度の高い記号列を行動として抽出する手法を提案した。実験結果より、平均再現率 86.0%、平均適合率 78.3% で人物の行動を抽出することができた。しかし、頻度だけの基準では、行動とまではいかない仕草や、行動の一部分などが、高い頻度で出現してしまう。そのため、抽出された部分記号列が行動を表しているかどうかを手動で判断することになり、ユーザーへの負担はまだ大きい。

今後の課題としては単位動作の持続時間の考慮など頻度以外の基準を設けることで、行動を表す記号列が上位になるように提示できると考えられる。また、記号列化せずに、単位動作の確率シーケンスに対して直接シーケンシャルデータマイニングなどで類似する部分シーケンスを発見する方法も、今後検討する予定である。

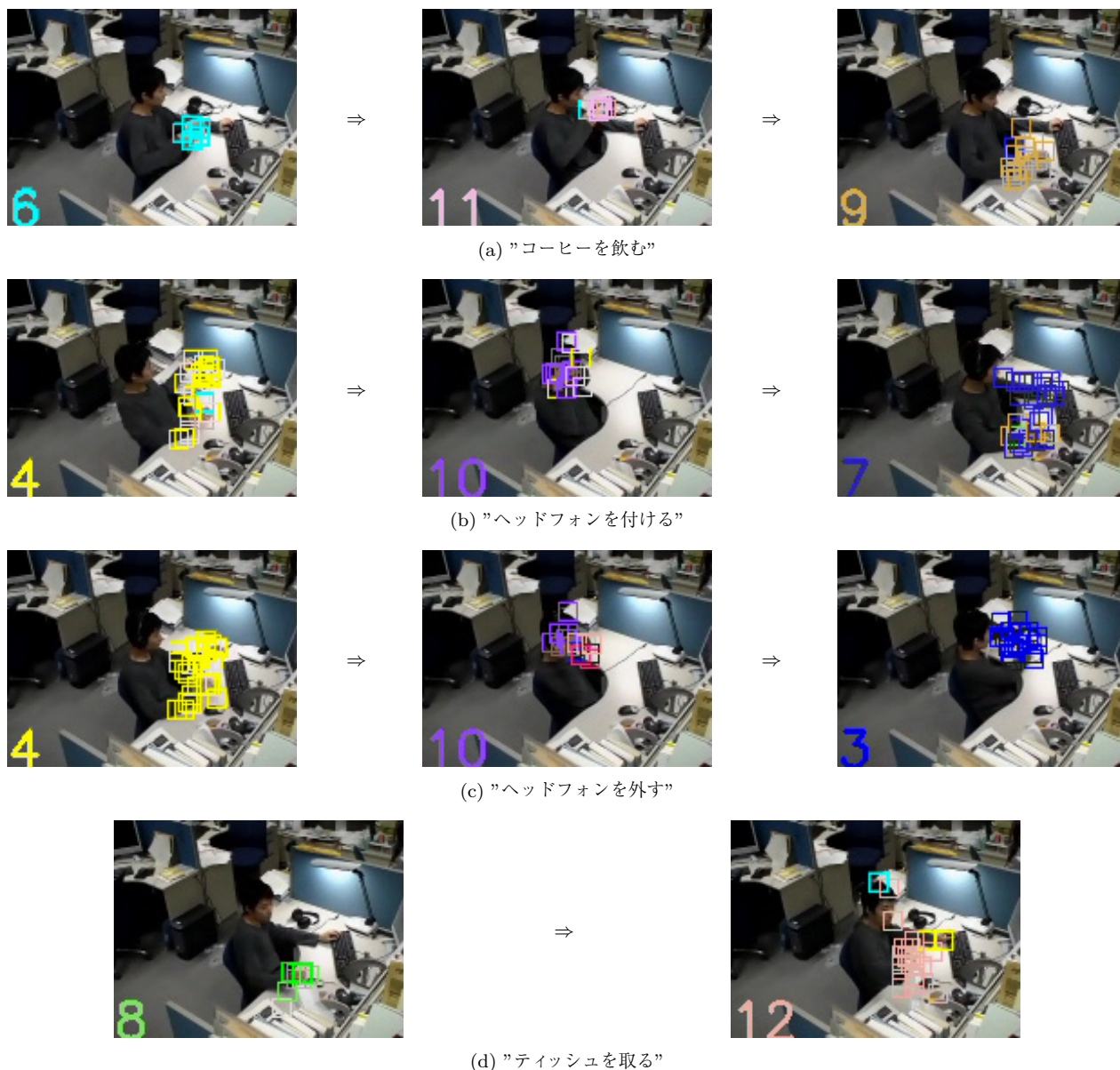


図 8 抽出された行動

文 献

- [1] T. Hofmann, "Probabilistic Latent Semantic Indexing," SIGIR, pp.50–57, 1999.
- [2] D. Gildea and T. Hofmann, "Topic-based Language Models Using EM," Proceedings of EuroSpeech, pp.2167–2170, 1999.
- [3] C. Schuldt, I. Laptev and B. Caputo, "Recognizing Human Actions: A Local SVM Approach," International Conference on Pattern Recognition, Vol.3, pp.32–36, 2004.
- [4] Yan Ke, Rahul Sukthankar, and Martial Hebert, "Efficient visual event detection using volumetric features," International Conference on Computer Vision, pp.166–173, 2005.
- [5] J.C. Niebles, H. Wang, Li.Fei-Fei, "Unsupervised Learning of Human Action Categories Using Spatial-Temporal Words," British Machine Vision Conference, Vol.3, pp.1249–1258, 2006.
- [6] J. Pei, J. Han, M. Behzad, and H. Pinto, "PrefixSpan: Mining Sequential Patterns Efficiently by Prefix-Projected Pattern Growth," International Conference on Data Engineering, pp.215–224, 2001.
- [7] Y. Ivanov and A. Bobick, "Recognition of Visual Activities and Interactions by Stochastic Parsing," IEEE Transactions on Pattern Analysis and Machine Intelligence, pp.852–872, 2000.
- [8] D. Moore and I. Essa, "Recognizing Multitasked Activities from Video Using Stochastic Context-Free Grammar," AAAI Conference on Artificial Intelligence, pp.770–776, 2002.
- [9] D. Minnen, I. Essa and T. Starner, "Expectation Grammars: Leveraging High-Level Expectations for Activity Recognition," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp.626–632, 2003.
- [10] R. Hamid, S. Maddi, A. Bobick, I. Essa, "Unsupervised Analysis of Activity Sequences Using Event-Motifs," Proceedings of the 4th ACM international workshop on Video surveillance and sensor networks, pp.71–78, 2006.
- [11] 木谷クリス真実, 佐藤洋一, 杉本晃宏, "映像に基づく人物行動の文法学習", MIRU, pp.149–155, 2007.