社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

人物動作の n-gram 表現に基づく物体認識

三木　博史†　　小島　篤博††　　黄瀬　浩一†

† 大阪府立大学大学院工学研究科　〒 599–8531 大阪府堺市中区学園町 1–1
†† 大阪府立大学総合教育研究機構
E-mail: †{miki@m.cs, kise@cs}.osakafu-u.ac.jp, ††ark@las.osakafu-u.ad.jp

あらまし　近年の物体認識に関する研究において，物体に対する人物の動作を間接的に利用することで，外観的特徴だけでは判別が難しい物体の認識が試みられている．これまでの人物動作に基づく物体認識では，人物動作のカテゴリを認識した後で，動作のカテゴリと物体とを関連付けるものがほとんどであった．本研究では，人物動作カテゴリは分類せずに、物体に関わるプリミティブな人物の動きを直接利用することで物体を認識する手法を提案する．まず，動画像から人物の動作特徴を抽出し，これを記号化して一連の動作を記号列で表現する．これらの記号列から n-gram を抽出し，人物動作を n-gram の集合として扱う．そして，物体カテゴリに関わる動作の n-gram 集合を用い，入力動画像から得られた人物動作と対応させることで物体の認識を行う．本報告では，物体ごとに行った動作から物体を認識する実験を行い，提案手法の有効性を確認した．
キーワード　環境認識，物体認識，人物動作認識, n-gram

Object Recognition Based on n-gram Representation of Human Action

Hiroshi MIKI†, Atsuhiro KOJIMA††, and Koichi KISE†

† Graduate School of Engineering, Osaka Prefecture University　1–1, Naka-ku, Sakai, Osaka, 599-8531 Japan
†† Faculty of Liberal Arts and Sciences, Osaka Prefecture University
E-mail: †{miki@m.cs, kise@cs}.osakafu-u.ac.jp, ††ark@las.osakafu-u.ad.jp

Abstract　In the field of object recognition, usage of human actions is applied to it, in order to tackle with the recongition of object that couldn't be identified by only using their appearance. Most of the conventional research for object recongition from actions, however, need an intermidiate classification of action category. In our research, primitive motions of human are directly apply to object recongition without classifying action categories. First, features of human motion are extracted from video images and encoded to symbols. Then, set of n-grams is obtained from the symbol sequence and registered to corresponding object category. At recognition phases, actions toward an object are also converted into n-grams which are to be compared with that registered in object categories. In this paper, we performed experiments of object recognition using our methods and confirmed the validity of this.
Key words　enviornment recognition, object recognition, human action recognition, n-gram

1. ま え が き

近年，オフィスや家の中などの屋内環境において，自律移動し種々のサービスを提供するロボットの研究開発が進められている．このようなロボットの要素技術の一つとして環境認識があり，特に物体のカテゴリ認識は，環境内の物体に関わる事象をロボットが扱うために重要であると考えられる．

これまでの物体認識を目的とした研究の多くは，物体の形状や色，テクスチャなどの外観情報を認識の手がかりとしてきた．しかしながら，人物が行動する室内環境は基本的に未知環境であり，オクルージョンや物が散乱しているなど，外観情報に基づく物体認識の障害となるものが多い．そのため，物体の外観情報だけでなく，物体と物体との関係や，物体と人物の動作との間の相互作用といった環境のコンテクストを利用した物体の認識が試みられている．特に人物動作を考慮することは，単に物体カテゴリを分類するだけでなく，物体の位置特定，その用途や機能といった，特定の環境内での物体の役割を含めた物体認識に有効であると考えられる．

そこで本研究では，物体に対する人物の動作に注目し，物体を推定する手法を提案する．人物動作の表現には，文書解析などで用いられる n-gram 表現を応用することで，プリミティブな動作を表現する．物体の推定においては，人物動作と物体と

の関係を自動的に学習させる．従来の手法が人物の動作カテゴリ認識を物体識別に応用しているのに対し，本手法では人物動作カテゴリ自体は認識せずに，特徴として得られる人物の動きから物体までを直接結びつけることで物体を推定する．

以下，2. では関連研究について述べる．次に，3. では本手法の概要について説明する．具体的な提案手法について，4. では人物の動作を特徴量としての抽出，5. では動作の特徴量から物体カテゴリを学習させる手法を説明する．本手法を用いて実験を行った結果を 6. に示し，7. と 8. で考察とまとめとする．

2. 関 連 研 究

物体認識の分野では，従来はその外観的特徴に基づく認識が試みられてきた [1]．こうした物体認識では，膨大な画像データから得られる特徴量に基づいて，物体カテゴリが学習されている．しかしながら，これらの手法では撮影角度による物体の見え方の変化や，オクルージョンの多い環境でのセグメンテーションに対応するのが難しく，見た目のみで物体を認識することは容易ではない．さらに，物体カテゴリを学習するためには，多数のモデルを用意する必要があり，その学習セットを作成すること自体容易なことではない．一方，物体の外観的特徴だけでなく，その物体を使う人物の行為・行動の観察を通して対象物の機能属性を認識し，その結果に基づいて対象物の認識を行うという新しい考え方が提案されている [2].

そこでまず，これまでの人物動作認識の関連研究について概説する．これまでの人物動作認識の多くは，バイオメトリクス，ビデオコンテンツの解析，セキュリティや監視システム，システム入力としてのジェスチャ認識などに応用されてきた．これらの人物動作解析の研究では，HMM や 3 次元ボリュームなどを利用した人物動作の特徴抽出，認識が試みられてきた．人物動作の取得にはカメラやモーションセンサなどの様々なデバイスから取得することが出来るが，ここでは一般的に用いられる映像情報を入力とした解析手法に注目する [3], [4]．映像による人物動作の解析には一般的に次のような手順がある．

（1） カメラからのビデオまたは画像列の入力
（2） 人物動作の特徴抽出
（3） プリミティブな動作認識
（4） 高レベルな動作認識

ここでは特に (2) 以降に注目する．画像列から抽出する人物動作の特徴として，オプティカルフローを用いる方法や，人物のシルエットに基づく特徴などが考えられている．また人物動作認識においては，(3) をどのように構成するかが重要となる．ここでのプリミティブな動作認識とは，特定のジェスチャ認識や，「座る」「立つ」といった基本的な動作の認識が含まれている．こうしたプリミティブな動作認識には，HMM，人体モデルへのフィッティング，人物シルエットのクラスタリングなどが用いられてきた．こうした基本的な動作の認識をベースとして人間の理解に近い高レベルな動作が認識されている．この (4) 高レベルな動作認識には，シーンの認識,「物を渡す」といった人物間での動作のインタラクションといった研究が含まれる．本研究のように，物体とそれに関わる人物動作とを関連付けた研究も，(4) の応用である．

こうした人物動作認識と物体認識を組み合わせた研究は，その双方を協調的に認識するものであった．Moore らは，モデルベースの物体認識と物体に対する人物の手の動きを相補的に用いることで，人物の行動と物体とを推定する手法を提案している [5]．また，樋口らは人物の動作と物体の機能や用途といった概念を階層モデル化することで，人物動作と物体の形状や機能の統合的な認識を試みている [6]．さらに物体と動作との関係を確率ネットワークで学習させる結び付けることで，より柔軟な認識を目指した手法も提案されている [7]．また，確率的に動作と物体との関係を学習させる手法で，物体の外観情報を用いることなく物体カテゴリを識別する研究もなされている [8]．これらの研究は，(3) プリミティブな動作認識をベースとして，(4) にあたる物体との関係を定義しようとするものであった．そのため，動作認識を主体とした場合の人物動作の分類及びカテゴリ分けが，そのまま物体認識に用いられてきた．しかしながら，物体に対する人物動作がどのように分類されるかは不明確であり，動作認識のための動作分類をそのまま物体認識に適用できるとは限らない．

一方，このようなカテゴリが不明な動作の認識に対して，文書解析で用いられてきた Bag-of-words アプローチを適用した研究がなされている．木谷らは，プリミティブ動作の教師なし学習のため，動き特徴に加えて物体の見えを考慮した特徴を用いている [9]．また，人物動作の時間的な系列を記述した特徴として，n-gram モデルを適用したものも興味深い．Hamid らは人物の行動を n-gram モデルで表現し，異常行動を監視する方法を提案している [10]．ただし，Hamid らの手法は (4) での高レベルな動作に対するアプローチであり，(3) でのプリミティブな動作の系列に対して n-gram モデルを適用したものである．また，(2) 人物動作の特徴に対して n-gram モデルを適用したものとしては，Thurau らによる研究があげられる [11]．この手法では，人物の姿勢の変化に対応するために n-gram モデルを適用している．こうした bag-of-words アプローチは，物体に対する人物動作を学習するのに適していると考えられる．それは，人物のプリミティブ動作と対象物体との対応が明確でないでも，動作を学習することができるからである．

そこで本研究では，特定的に動作カテゴリを学習する HMM のような方法ではなく，(2) での人物動作の特徴から直接的に物体を推定する手法を提案する．そのために必要な人物動作の特徴には，n-gram を用い，人物動作の時間的変化を考慮したものを作成する．

3. 提案手法の概要

本手法では，物体に対して観測される人物の動作の集合を調べることで，物体カテゴリの表現，推定を行う．このとき，人物の動作カテゴリの認識は行わず，直接人物の動きと物体カテゴリとを関連付ける．本手法の前提条件として，推定対象となる物体候補，および物体に対してなされた人物動作は検出可能であるとする．また，人物の動作取得にはステレオカメラを用い，視覚的特徴だけでなく 3 次元位置情報を用いる．

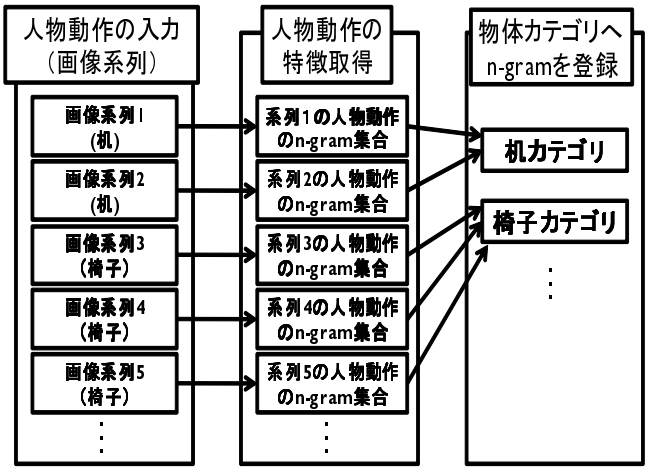


図 1　物体カテゴリに対する人物動作の学習手順

(a) 入力画像　　(b) 肌色領域

図 2　肌領域抽出

図 1 に人物動作から物体カテゴリを学習する手順を示す．まずステレオカメラを用いて物体に対する人物の動作を撮影する．図 1 での各画像系列は，時間的に連続した区間で人物が物体を扱った一連の動作を記録したものである．例えば，棚に対する人物動作であれば，人物が棚の扉を開け，物を取り出し，扉を閉めて棚から離れるまでの一連の動作を一つの画像系列とする．次に，撮影された画像系列から，人物動作を表す記号列を抽出する．これらの記号列から n-gram を作成し，一連の動作に含まれるプリミティブ動作の集合を抽出する．抽出された n-gram を物体カテゴリに登録することで，物体カテゴリを学習させる．認識処理においては，認識したい物体に対する人物動作の画像シーケンスから，学習時と同様に人物動作を表す n-gram 集合を抽出する．そして，抽出した n-gram と物体カテゴリに登録された n-gram とを比較することで，物体を認識する．

4.　人物動作の特徴抽出

本研究では, 顔と手の動きを追跡することで, 人物の動作特徴を抽出する．顔と手に注目する理由の一つは, 顔の位置を追跡することで, 立っている, 座っているなどの人物の姿勢が得られることである．また, 物体と関連する人物動作の多くが手を使う動作であるため, 手の位置を追跡することでそれらの手を使う動作を抽出できると考えられる．

4.1　顔領域および手領域の追跡

顔と手の動きを追跡するため, まず画像中の肌色情報を抽出する．肌色情報抽出に関する研究としては, 画像中の肌色情報を抽出し, それらの領域を追跡するものがある [12]．これは, 肌色の確率分布を肌色モデルとして作成し, 画像中の肌確率を求めることで肌色領域を抽出する手法である．本手法では, $L^*a^*b^*$ 表色系のうち, a^*, b^* を用いて肌色モデルを作成し, 図 2(a) のような入力画像から, 図 2(b) のような肌領域を抽出する．

本手法では, 人物の動作を観察する際にはカメラの位置は固定されているものとし, 肌色領域の 3 次元位置をステレオカメラによって取得する．得られる肌色領域のうち, 最も高い位置にあるものを顔領域とする．その他の領域で, 顔からの 3 次元距離が一定範囲内のもので, 最も領域サイズの大きいものを手領域とする．

表 1　動 作 特 徴

動作特徴	ラベル	量子化数
顔の高さ	H_f	3
手と顔の距離	D_{fh}	3
手から顔への方向	A_{fh}	5
手と顔の距離変化	R_{fh}	3
顔の動く速さ	S_f	3
手の動く速さ	S_h	3
顔の動く方向	M_f	6
手の動く方向	M_h	6

4.2　人物の動作特徴

前節の処理で得られた顔と手の 3 次元位置から, 人物が物体を扱う際の動作の特徴量を抽出する．本手法では, 表 1 に示される動作特徴を, 特徴量として考える．これらの動作特徴には, 単一のフレームから算出できるもの $(H_f, D_{fh}, A_{fh},)$ と, 現在のフレームと直前の T フレームにおける動作から算出されるもの $(R_{fh}, S_f, S_h, M_f, M_h)$ とがある．カメラ座標系で, 水平方向を x, 鉛直上向きを y, 奥行きを z とし, 時刻 t での顔と手の位置を $F(t) = (f_x(t), f_y(t), f_z(t))$, $H(t) = (h_x(t), h_y(t), h_z(t))$ として, 各動作特徴を次式のように定める．

$$H_f(t) = f_y(t) \tag{1a}$$

$$D_{fh}(t) = |F(t) - H(t)| \tag{1b}$$

$$A_{fh}(t) = \arccos \frac{f_y(t) - h_y(t)}{D_{fh}(t)} \tag{1c}$$

$$R_{fh}(t) = D_{fh} - \frac{1}{T} \sum_{k=t-T}^{t-1} D_{fh}(k) \tag{1d}$$

$$S_f(t) = |F(t) - F^{'}(t)| \tag{1e}$$

$$S_h(t) = |H(t) - H^{'}(t)| \tag{1f}$$

$$M_f(t) = \arccos \frac{f_y^{'}(t) - f_y(t)}{S_f(t)} \tag{1g}$$

$$M_h(t) = \arccos \frac{h_y^{'}(t) - h_y(t)}{S_h(t)} \tag{1h}$$

ただし,

$$F^{'}(t) = \frac{1}{T} \sum_{k=t-T}^{t-1} F(k), \qquad H^{'}(t) = \frac{1}{T} \sum_{k=t-T}^{t-1} H(k)$$

ここで, A_{fh}, M_f, M_h は, 鉛直下向きからの角度である．すなわち, 水平方向への動作であれば, これらの角度は $\pi/2$ で

ある．

次に，得られた動作特徴を量子化する．これは，各時刻における人物の動作を離散的に表現し，記号化するためである．

- 顔の高さ (H_f)

人物が立っている状態，椅子に座っている状態，しゃがんでいる状態の 3 状態を区別するための閾値を設定する．立っている状態と座っている状態での顔の高さの平均値，また同様に，座っている状態としゃがんでいる状態での顔の高さの平均値を閾値とする．

- 顔と手の動く速さ（S_f, S_h)

現在のフレームより前の数フレームと比較して，静止している，ゆっくり動いている，速く動いている，3 状態を区別できる閾値を設定する．ゆっくり動く速さは，物をとるときの手の動く速さを基準とする． また，速く動く場合の速さは，歩くときの顔の動く速さを基準とする．

- 顔，手の動く方向，手から顔への方向 (M_f, M_h, A_{fh})

鉛直下向きからの角度を 5 段階に量子化したものとなる．このときの閾値は，$\frac{\pi}{8}$, $\frac{3\pi}{8}$, $\frac{5\pi}{8}$, $\frac{7\pi}{8}$ である．また，顔，手の動く方向については，静止している状態を含め 6 段階である．

- 顔と手の距離 (D_{fh})

顔と手の距離が近い状態，肘を曲げた状態，肘を伸ばした状態の 3 種類が区別できるように閾値を設定する．

- 顔と手の距離の変化 (R_{fh})

顔と手の距離が，変わらない，近づく，遠ざかる，の 3 状態を区別できるように閾値を設定する．

以上のように動作特徴を量子化し，次のように各フレームでの人物の動きを記号 x_t で表し，動作記号と呼ぶ．

$$x_t = (H_f, D_{fh}, A_{fh}, R_{fh}, S_f, S_h, M_f, M_h) \tag{2}$$

5. 物体カテゴリに対する人物動作の学習

前章で説明した人物動作特徴の記号系列 $(x_0, x_1, \ldots, x_t, \ldots)$ から，物体カテゴリと関わる人物の動きを学習する．本手法では，人物動作特徴の記号系列から n-gram を作成し，物体カテゴリへ登録することで学習を行う．これは，4. で述べた特徴記号をそのまま物体カテゴリへ登録するだけではその時間的変化を考慮することができないためである．

5.1 人物動作の n-gram 表現

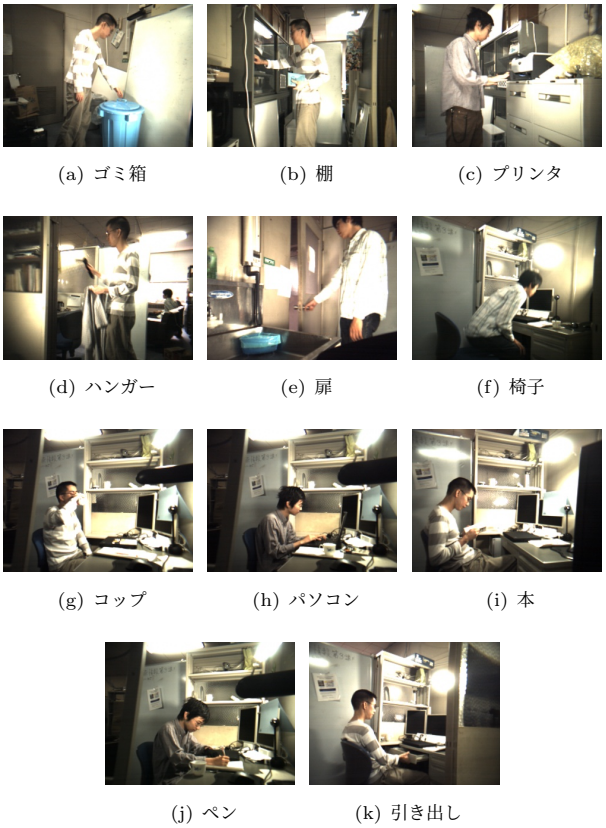
4. で説明した単独の動作記号 x_t に加え，連続する n フレーム分の記号系列を考慮することで，時間的変化に対応する人物動作を表した n-gram w_t^n を作成する．人物動作特徴の記号系列 $(x_0, x_1, \ldots, x_t, \ldots)$ に対して，n-gram w_t^n を次のように定める．

$$w_t^n = \boldsymbol{x}_t \boldsymbol{x}_{t+1} \ldots \boldsymbol{x}_{t+n-1} \tag{3}$$

このようにして得られた n-gram $w_t^n (n = 1, 2, \ldots, L)$ を物体カテゴリへ登録する．ただし，L はここで考慮するフレーム数の最大値を表している．

5.2 物体カテゴリへの特徴系列の登録

前節で得られた w_t^n の集合で物体カテゴリを表現する．物体カテゴリ o に対する人物動作 w_t^n の集合を S_o とする．次に，各物体カテゴリ間で重複する n-gram を除去する．これは，座っている，立っているといった，人物の姿勢に大きく影響を受ける動作特徴を除去し，さらに異なる物体カテゴリに重複して現れる動作を除去することで物体識別に有効な動作集合のみを残すためである．



(a) ゴミ箱　(b) 棚　(c) プリンタ

(d) ハンガー　(e) 扉　(f) 椅子

(g) コップ　(h) パソコン　(i) 本

(j) ペン　(k) 引き出し

図 3　11 種類の物体と物体に対する動作

カテゴリの推定に用いる新たな人物動作の集合を S'_o とし，次のように定める．

$$S'_o = S_o - \bigcup_{i \neq o} S_i \tag{4}$$

このようにすることで，物体 o に対する人物動作の n-gram のうち，物体 o のみに対するものだけが残される．

5.3 物体カテゴリの認識

最後に，物体カテゴリの認識処理について述べる．推定すべき物体 Q に対する人物動作系列が与えられたとき，5.1 と同様にして，推定対象物体に対する動作集合 S_Q を求める．次に，S_Q と学習させた物体 O の動作集合 S'_O との積集合を求め，その要素数を物体カテゴリのスコアとする．すなわち，

$$score(Q, O) = |S_Q \cap S'_O| \tag{5}$$

このスコアを全ての物体カテゴリについて求め，最大のスコアを得た物体カテゴリを物体 Q の認識結果とする．

6. 実　　験

提案手法の有効性を確認するため，物体カテゴリを識別する

表 2　テストデータに学習時と同じ人物が含まれる場合の認識結果.

	認識率 (%)										
認識対象	A	B	C	D	E	F	G	H	I	J	K
A: ゴミ箱	93	1	0	1	3	0	0	0	0	0	0
B: 棚	4	83	6	7	0	0	0	0	0	0	0
C: プリンタ	7	36	30	22	0	1	0	0	0	0	0
D: ハンガー	0	20	8	71	0	0	0	0	0	0	0
E: 扉	12	3	0	0	83	0	0	0	0	0	0
F: 椅子	1	0	0	0	0	94	1	0	0	0	0
G: コップ	1	0	0	0	0	0	62	7	16	12	2
H: パソコン	0	0	0	0	0	0	6	50	27	16	1
I: 本	0	0	0	0	0	0	0	10	81	7	2
J: ペン	0	0	0	0	0	0	3	10	12	71	3
K: 引き出し	0	0	0	0	0	1	4	1	17	7	70

実験を行った．対象物体は 11 種類とした．一つの物体カテゴリに対して 3 人の人物が 30 回ずつ動作を行い，合計 90 シーケンス撮影した．個々のシーケンスでは物体に対する一連の動作すべてが含まれている．たとえば棚に対する動作では，開ける，物を取り出す（しまう），閉めるといった動作のすべてが含まれている．これらの 11 物体に対する動作を撮影したものを図 3 に示す．これらの物体のうち，(a)〜(e) は立った状態で扱う物体，(f) は立つ状態と座る状態が変化する動作を受ける物体，(g)〜(k) は座った状態で扱う物体である．各物体カテゴリの撮影で対象とした物体は一つである．また，撮像条件を変化させるため，一つの物体につき異なる 3 方向から撮影した．物体の撮影には SRI Internationl 社製ステレオカメラを用いた．画像の解像度は 320 × 240 ピクセル，ステレオカメラによる奥行値計算には，ステレオカメラに付属のライブラリを使用した．撮影のフレームレートは 15fps で行った．

6.1　実験 1：テストデータに学習時と同じ人物が含まれる場合

まず，各物体で 20×3 人分 =60 シーケンスを学習データとし，残りの 30 シーケンスをテストデータとする．全てのシーケンスがテストできるように学習データとテストデータを入れ替え，90 シーケンスすべての認識を行った．このとき，学習データとテストデータには，同じ人物が行った動作がそれぞれ含まれている．また，この実験では動作特徴抽出での $T = 5$, n-gram の最大長を $L = 10$ とした．

上記の条件で認識実験を行った結果を表 2 に示す．表 2 の対角成分がカテゴリを正しく認識した割合である．この実験では，プリンタ以外の物体では正解物体の認識率がもっとも高く，正しく認識することができている．また，立った状態で扱う物体 (物体 A〜E) と座った状態で扱う物体 (物体 G〜K) で混同して認識されることはなかった．

6.2　実験 2：学習と認識とで対象人物が異なる場合

次に，テストデータで物体に対する動作を行った人物が学習データに含まれていないという条件のもとで実験した．すなわち，各物体で 30×2 人分 =40 シーケンスを学習データとし，残りの 1 人分 30 シーケンスをテストデータとした．実験 1 と同様に，90 シーケンス全てがテストできるように学習データとテストデータを入れ替えて実験した．この実験でも $T = 5$,$L = 10$ とした．

表 3　テストデータと学習時で同じ人物が含まれない場合の認識結果

	認識率 (%)										
認識対象	A	B	C	D	E	F	G	H	I	J	K
A: ゴミ箱	88	2	0	0	9	0	0	0	0	0	0
B: 棚	3	70	11	16	0	0	0	0	0	0	0
C: プリンタ	3	50	22	21	0	2	0	0	0	0	0
D: ハンガー	1	40	14	43	0	0	0	0	0	0	0
E: 扉	16	2	0	0	82	0	0	0	0	0	0
F: 椅子	0	0	0	0	0	94	1	0	0	0	0
G: コップ	0	0	0	0	0	0	54	4	23	10	8
H: パソコン	0	0	0	0	0	0	7	27	42	19	4
I: 本	0	0	0	0	0	0	2	10	78	7	3
J: ペン	0	0	0	0	0	1	4	12	31	43	8
K: 引き出し	0	0	0	0	0	1	8	1	23	9	58

表 4　記号長 L を変化させたときの正答率 (%)

	記号長 L					
	1	3	5	7	10	100
人物 1	25	52	59	65	66	65
人物 2	37	48	52	53	53	53
人物 3	28	52	55	61	61	61

その認識結果を表 3 に示す．表 2 と比べると，表 3 でも立った状態と座った状態で扱う物体とでの認識が混同しておらず，認識回数の分布は同様の分布を示している．この結果，別人の学習データを用いて物体を認識した場合でも，本手法では正しく認識することが可能であると言える．しかしながら，正しく認識出来た回数は全体的に低下している．また，プリンタに加えて，パソコンを正解確率が大幅に低下し，別の物体であると認識された確率のほうが正解確率より大きくなった．

6.3　実験 3：記号長 L を変化させたときの認識率

この実験では，最大記号長 L を変化させることによって，人物動作を表す n-gram を作成する本手法の有効性を確認する．実験条件は実験 2 とほぼ同じであり，最大記号長 L のみを変化させて 3 人の人物ごとでの正答率を求めた．ここで，記号長 $L = 1$ の場合は，5.1 で述べた n-gram を考慮せずに，4.2 での人物の動作特徴のみを用いて認識することを意味している．

実験の結果を表 4 に示す．n-gram を全く考慮しない場合 ($L = 1$) に比べて，最大記号長 $L = 10$ とした場合では認識率が 2 倍程度になっており，n-gram を考慮した本手法が有効であることが分かった．しかしながら，L を 7 以上としたとき，正答率にほとんど変化はなく，ある程度の長さ以上の n-gram は認識にはほとんど影響を与えないことが分かった．

7.　考　　察

6.1 での実験結果から，人物動作カテゴリに注目することなく物体を認識することは十分可能であることがわかった．また，6.2 での結果からは，学習時とは別人の動作を利用する場合でも物体識別を行うのは可能であるとの結論を得た．これらの実験に対して，いくつかの項目について考察する．

7.1 人物ごとに異なる動作の影響

6.1 と 6.2 との比較では，6.2 での認識結果は全体的に悪化しており，異なった人物の動作を用いた場合の影響が問題となっている．こうした問題の対策としてまず考えられるのは，学習サンプルに用いる人物数を増やすことである．今回の実験では 3 人分の実験データのみを用いているため，人物の癖などを含んだ動作の影響を吸収できる学習データを作成するにはデータセットは十分とは言えない．表 4 での結果でも人物ごとに正答率が大きく異なっており，これらの正答率が人物ごとに大きな差が出ない程度に，サンプルに用いる人数を増やす必要がある．

7.2 認識の難しい物体について

6.2 では，特にプリンタ，パソコン，ペンなどで正答率が低かった．そうした物体に対する動作は，変化が少ない，もしくは一つの動作にかかる時間が短いものが多かった．たとえばプリンタに対する動作では，ほとんどの場合紙をとる動作だけであり，さらにプリンタ自体に触れる時間は極めて短い．また，パソコン，ペンなどは，座った状態であまり動かない動作であるため，動作の変化が少ない．こうした動作は，他の物体カテゴリに対する複雑な動作の一部になっていることが考えられる．こうした物体に対する動作は，5.2 で述べたように，他カテゴリと被る n-gram のすべてを排除していることによって，認識に有効な n-gram も排除されていると考えられる．そのため，物体の認識により有効な n-gram を残す方法をさらに検討する必要がある．

7.3 人物の動作を取得する方法の妥当性

本手法では，4. で述べたように，人物の動作から得られる実数値の特徴を量子化している．実験結果から，この量子化によって，人物の動作を表現する適切な n-gram を生成できていることを確認した．しかしながら，この量子化を細かくしすぎた場合，人物の動作を n-gram に変換したときに，その n-gram が他の同様の動作から得られる n-gram と一致する可能性は低くなると考えられる．そのため，表 1 での量子化数は注意深く決定する必要がある．この量子化数についての検証はまだ不十分であり，どの程度の量子化数が適当かは今後検討していかなければならない．

また前節でも述べたように，人物動作を表現した n-gram を物体カテゴリにどのように登録していくかは今後検討する必要がある．本手法では n-gram の頻度を全く考慮していないが，n-gram の頻度による重みづけをすることで，検索に本当に有効な n-gram を検出するといったことが考えられる．

8. ま と め

本研究では，物体の外観情報を用いることなく，人物動作から物体カテゴリを推定する手法を提案した．本手法では，人物動作カテゴリを考慮することなく，人物動作の n-gram を作成することによって物体に対する人物動作を表現し，物体カテゴリと人物動作を結び付けた．実験では，人物動作のカテゴリを考えることなく，人物動作を物体認識に適用する本手法の有効性を示した．

文　献

[1] 柳井：“一般物体認識の現状と今後”, 情報処理学会論文誌: コンピュータビジョン・イメージメディア, **48**, SIG 16, pp. 1–24 (2007).

[2] M. Kitahashi, A. Kojima, M. Higuchi and K. Fukunaga: “Toward a cooperative recognition of human behaviors and related objects”, 15th European Japanese Conference on Information Modelling and Knowledge Bases, pp. 321–330 (2005).

[3] T. Moeslund, A. Hilton and V. Kruger: “A survey of advances in vision-based human motion capture and analysis”, Computer vision and image understanding, **104**, 2-3, pp. 90–126 (2006).

[4] P. Turaga, R. Chellappa, V. Subrahmanian and O. Udrea: “Machine recognition of human activities: A survey”, IEEE Transactions on Circuits and Systems for Video Technology, **18**, 11, pp. 1473–1488 (2008).

[5] D. Moore, I. Essa and M. Hayes: “Exploiting human actions and object context for recognition tasks”, IEEE International Conference on Computer Vision 1999 (1999).

[6] M. Higuchi, S. Aoki, A. Kojima and K. Fukunaga: “Scene recognition based on relationship between human actions and objects”, 17th International Conference on Pattern Recognition, Vol. 3, pp. 73–78 (2004).

[7] A. Gupta and L. S. Davis: “Objects in action: An approach for combining action understanding and object perception”, Proc. IEEE Conference on Computer Vision and Pattern Recognition CVPR ’07, pp. 1–8 (2007).

[8] 三木, 小島, 宮本, 黄瀬：“DBN を用いた人物の動作パターンに基づく物体認識”, 画像の認識・理解シンポジウム (MIRU2008) 論文集, IS3–3, pp. 877–884 (2008).

[9] 木谷 クリス, 岡部, 佐藤, 杉本：“視覚的文脈を考慮した人物動作カテゴリの教師なし学習”, 画像の認識・理解シンポジウム (MIRU2008) 論文集, OS1–4, pp. 28–33 (2008).

[10] R. Hamid, A. Johnson, S. Batta, A. Bobick, C. Isbell and G. Coleman: “Detection and explanation of anomalous activities: Representing activities as bags of event n-grams”, IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005. CVPR 2005, Vol. 1 (2005).

[11] C. Thurau and V. Hlavac: “Pose primitive based human action recognition in videos or still images”, IEEE Conference on Computer Vision and Pattern Recognition, 2008. CVPR 2008, pp. 1–8 (2008).

[12] A. Kojima, T. Tamura and K. Fukunaga: “Textual description of human activities by tracking head and hand motions”, 16th International Conference on Pattern Recognition, Vol. 2, pp. 1073–1077 (2002).