身体各部の動作概念の階層性に基づく 動画像中の人物行動のテキスト表現

正員 小島 篤 博 (大阪府立大学総合情報センター)

非会員 浅沼 克 紀 (大阪府立大学工学部)

正員 福永 邦 雄 (大阪府立大学工学部)

Text Generation about Human Behavior from Video Images Using Hierarchical Concepts of Motions of Body Parts

Atsuhiko Kojima, Member (Library and Science Information Center, Osaka Prefecture University),
Katsunori Asanuma, Non-member, Kunio Fukunaga, Member (College of Engineering, Osaka
Prefecture University)

In this paper, we propose a method of generating text about human behavior from video images using hierarchical concepts of motions of body parts including body, head and hands. In natural language, there are various expressions about human acts, so we firstly analyze and classify concepts of several motion verbs about each body part using case frame expressions which represent semantic structure of natural language sentences. By making correspondence between these concepts and estimated positions, postures and motions of human body parts from video images, appropriate verbs, objects, locus etc. are determined and expressed in case frames. Lastly these case frames of each body part are integrated into a total body expression, and translated into natural language text by applying case patterns and syntactic rules.

キーワード：自然言語生成, 階層的動作概念, 格フレーム, 画像理解, 動作推定, 物体への作用

1. ま え が き

近年, さまざまな形態の画像から自然言語のテキストを生成する研究が行われている⁽¹⁾⁽²⁾。とくに人物や動物体の動きをテキストで表現する研究としては, 自動車などの乗物の移動を追跡し, その動きを適切に表現する動詞を生成する研究⁽³⁾⁽⁴⁾, 人物の位置や移動速度をもとに事象を抽出し, あらかじめ与えられたシナリオに沿って一連の事象を解釈する研究⁽⁵⁾などが挙げられる。

しかしながら, これら映像による監視の研究では, 映像から得られた動物体の位置・姿勢などが特定の条件を満たしているかどうかを監視することに主眼が置かれ, それを報知・確認するための手段として比較的簡易な処理でテキストを生成しているのが現状である。

一方, 自然言語の事象概念は多様であり, とくに人物の動作に関する概念には豊富なバリエーションがあり, 同じ動作を認識したとしても, それを行った場所や対象によって表現は様々に変化する。この問題を扱うためには, 動詞の意味を時間的・空間的な概念に基づいて分析するなど, 自然言語分野における知見の導入が不可欠である。このよう

な研究としては, 動画像ではないものの線画情報から概念的な特徴を抽出し, 意味解釈を行った上でテキストとして出力する研究が挙げられる⁽⁶⁾⁽⁷⁾。また動画像を対象にしたものとしては, 人物の頭部に注目して位置・姿勢推定を行い, さらに画像から得られる数値情報からさまざまな特徴を自然言語の事象概念の成分として組み立てることでテキストを生成する研究⁽⁸⁾が試みられている。

しかしながらこれらの研究においては, 図中の対象や人物頭部の位置など単一の情報をもとに事象を解釈しているため, 手の動きや物体への作用など身体各部が連携するような複雑な動作を適切に表現できない場合があるなどの問題があった。とくに物体への作用は, その物体の持つ属性などによって表現が多様に変化するため, これを統合的に扱う枠組みが必要である。

本研究では, 身体の位置・姿勢, 頭部の向き, 手の位置について動作の概念を階層的に分類し, 画像から得られた情報をもとにそれぞれの固有の動作表現を生成した上で, 動詞の特性に応じた規則に基づいて統合し, 最終的に自然言語のテキストを生成する手法を提案する。

2. 動画像からのテキスト生成

ここでは、動画像からテキストの生成までの全体的な処理の概要を、図 1 にしたがって説明する。まず人物の行動を画像から推定する必要があるが、人体のような複雑な多関節物体の位置や姿勢を実時間で推定するのは現状では困難である。そこで本手法では、比較的軽い処理で身体各部の動作を効果的に得るためには、以下の3つを推定すれば、オフィス環境など室内での動作を推定することを目的とした場合、ほぼ十分な情報を得られると考えた。

頭部の位置 身体と、立っているか座っているかといった姿勢が得られる。

頭部の姿勢 どの方向を向いているか、さらにはどの物体を見ているかといった視線情報が得られる。

手の位置 触る、操作するなど対象物への作用が得られる。さらに、物体への作用については人物の身体の情報に加え、対象物体をシーン中に“置く”、“取る”といった動作の推定、またその物体の識別なども行う必要がある。

本手法ではまず、色情報を利用して肌領域を抽出し、人物頭部と手の位置を座標変換により求める。頭部の姿勢はあらかじめ様々なアスペクトのモデルデータを用意しておき、入力画像の頭部領域と比較することで推定する。また対象物体に対する動作は、背景との差分により得られる物体領域の形状と入力画像のエッジを比較することにより判別し、同時に物体の識別も行う。

次に、このような画像から得られた情報と、あらかじめ与えられた環境の知識を用いて、身体各部の動作表現を生成する。自然言語の動詞や名詞などが表す概念は、成分的な特徴である意味素性の組み合わせにより構成されており、意味素性が少ないほど抽象的な表現となり、また多いほど具体的で詳細な表現となる⁽⁶⁾。そこで本手法では、人物の身体の一部ごとに動作や状態をとらえ、画像情報から抽出可能な意味素性と照合することで段階的に動作概念を詳細化し、部位に関する動作を表現した格フレームを生成する。ここで用いる格フレームとは、自然言語処理において文の意味構造を表すために利用される表現形式である。

最後にこれらを統合することで人物の総体的な動作・状態を表す格フレームを生成し、格構造パターンと構文規則を適用することで自然言語のテキストに変換する。

3. 身体各部の動作・物体への作用の抽出

本章では、人物頭部の位置・姿勢、手の位置ならびに物体への作用を画像から抽出する方法について述べる。

〈3・1〉 頭部の位置・姿勢および手の位置 動画像において人物の顔や手などを追跡する手法はジェスチャ認識を中心にいくつか提案されているが、本手法では複雑な背景環境においてもロバストに肌領域を追跡することができる文献(9)の手法を利用している。この手法では、人物がいない状態で一定時間撮影した背景画像と、あらかじめ複数の人物の画像から取得した肌色の色域から、入力画像の各

図 1 処理の概要

Fig. 1. Outline of the process.

画素ごとに背景らしさと肌色らしさを確率的に推定する。さらに、過去数フレーム分の肌領域の移動から各画素が肌領域に含まれる確率を推定する。そして、これら3つの確率を Dempster-Shafer 理論⁽¹⁰⁾により統合することで入力画像の各画素が肌領域である確率を求め、閾値処理により肌領域を抽出する。

今回の実験では画像中の人物は一人と仮定しているため、得られる肌領域は最大で顔と両手の3つであり、最も面積が大きくかつ高い位置にある領域を顔の領域とした。また、画素が背景である確率が小さい領域をもとに人物領域を得ることができ、このうち顔領域の周辺を頭髮領域として含め、頭部領域とする。

次に抽出した頭部および手の3次元位置推定を行う。まず、本手法で用いる座標系を図2に示す。O-XYは画像の中心を原点とする画像平面上の座標系である。また、O_c-x_cy_cz_cはカメラ座標系であり、視点の位置を原点、視線の方向をz_c軸、そして $t \equiv [t_x \ t_y \ t_z]^T$ をカメラ座標系における人物頭部の位置とする。

カメラ座標系上の点 $[x_c \ y_c \ z_c]^T$ は、奥行きz_cが定まれば、画像平面上の位置X,Yから透視射影による座標変換により次式を用いて求めることができる。

$$\begin{bmatrix} x_c \\ y_c \end{bmatrix} = \frac{d}{f} \cdot z_c \begin{bmatrix} X \\ Y \end{bmatrix} \dots\dots\dots (1)$$

ここで、dは画素の大きさ、fはカメラの焦点距離を表し、これらの値は既知量とする。人物頭部の大きさを既知とし、

図2 位置推定における座標系
Fig. 2. Coordinate systems for position estimation.

図3 頭部の姿勢推定
Fig. 3. Posture estimation of human head.

画像平面上の大きさから奥行き t_z を求めて $z_c = t_z$ として (1) に代入すれば、人物頭部の位置 t が得られる。同様に、手の位置についても奥行きが頭部と同じと仮定することで求めることができる。

次に頭部の姿勢推定は、図3に示すように入力画像から切り出した頭部領域に対し、あらかじめ撮影しておいた様々な姿勢のモデルとマッチングをすることで推定する。まず、切り出した頭部領域をモデル画像と同じスケールにリサイズし、これを I とする。頭部の俯角 ϕ 、水平角 θ の回転に対応するモデルを $M_{\phi, \theta}$ として、入力 I とモデル $M_{\phi, \theta}$ の類似度を各画素の色の相関により求める。そして、入力と全てのモデルとの類似度を求め、最も評価値の高いモデルの姿勢 ϕ, θ を入力の姿勢とする。なお、本手法では ϕ 3 方向、 θ 8 方向の計 24 種類のモデルを用いている。

〈3・2〉対象物への作用 人物の行動の多くは、対象物との相互作用に基づいている。本手法では、対象物に対して“取る”、“置く”などの作用を及ぼす動作の瞬間に注目し、対象物の位置および種類を認識する。入力画像に対し〈3・1〉で用いた画素ごとの背景らしさの確率を計算すると、背景に対する変化領域を得ることができる。画像中に存在する人物が1人であり他の動物体が存在しないと仮定すると、変化領域画像中に人物領域以外の孤立領域が発生した場合には、物体をシーン中に新たに配置したか、またはシーン中の物体を取り除いたと考えることができる。

(a) Difference image in which a man is picking (left) or putting (right) an object.

(b) Input image for each image above. Man and object regions are circumscribed in rectangles.

図4 物体領域が出現した瞬間の変化領域
Fig. 4. Difference regions when object region appears.

図5 物体領域のエッジ
Fig. 5. Edge of object region when object is picked or put.

図4(a)は物体領域が出現した時の変化領域画像を示しているが、これだけでは物体を“取った”のか“置いた”のかを判断するのは難しい。図4(b)はこれらの変化領域に対応する領域を入力画像上で矩形で囲ったものであるが、“取った”場合は入力画像上の対応領域には物体が存在せず、“置いた”場合は同領域に物体が存在している。そこでこの入力画像と変化領域画像のそれぞれでエッジを抽出してみると、図5に示すように、物体を取った場合と置いた場合では、両者の相関に明らかに違いが見られる。

そこでこれを評価するために、物体領域において入力画像と変化領域画像におけるエッジの類似度をそれぞれのエッジ点の重なる割合と定義し、これが閾値より大きければ、入力画像と変化領域画像の双方において同じ物体のエッジが存在するとして、物体を“置いた”と判断する。

なお本手法では、この物体領域の背景情報を逐次更新しているため、一度環境中に置かれた物体は一定時間経過すると背景として扱われる。

〈3・3〉形状と色を用いた物体の識別 前節の処理において動作の判別を行う際、同時に物体領域が得られているため、この領域に対してあらかじめ想定した物体のモデルとのマッチングを取ることで物体の識別を行う。本手法では室内中の人物の動作対象としてノート PC、コップ、ラジカセなど 9 種類の物体を想定し、物体領域とモデルとの比較に形状と色の 2 つを利用する。

まず、入力画像上での物体領域から物体の形状情報としてエッジを抽出する。同様にモデル物体に対してもエッジ抽出を行うが、物体の形状はそのアスペクトにより変化するため、本手法では各モデル物体につき 4～8 種類のアスペクトを用いている。次に、モデルのエッジ画像を、入力エッジ画像に合わせて縦横比が変わらないようにリサイズした上で、両者の比較を行う。ここで、入力とモデルのエッジ画像を直接比較すると、エッジ点のズレによって類似度が大きく影響を受け、正確な物体の識別が困難となる。そこで、このようなエッジ点同士のズレの影響を回避し大局的な類似度を評価するために、それぞれのエッジ画像に対してガウシアンフィルタを適用する。そしてこれらの画像間で相関をとり、これを入力の物体とモデル物体の形状情報の類似度とする。

次に、物体の色情報を用いた物体のマッチング方法について述べる。まず、入力画像の物体領域中の各画素を、明度変化の影響を考慮して CIE1976L*a*b* 表色系⁽¹¹⁾に変換し、このうち a*, b* による単位色域上のヒストグラムを作成する。同様に、モデル物体のサンプル画像からヒストグラムを作成する。そして、それぞれのヒストグラムを全画素数で割り、[0, 1] に正規化した後それぞれの相関をとり、類似度を算出する。

このようにして、入力画像中の物体に対して、各モデル物体との形状情報および色情報の類似度を求め、これらの積を総合的な類似度として最も評価の高いモデル物体を入力画像中の物体とする。

4. 動作概念による人物行動の認識

一般に画像から得られる人物の 3 次元位置やその時系列上の変化といったパターン情報と、自然言語のテキストのような概念的・記号的な情報との間には大きなギャップがあり、両者の対応づけは難しい。そこで本手法では、まず画像から得られる身体各部の位置・姿勢の数値データ列から抽出可能な動作特徴と、自然言語における事象概念、特に人物の動作に関わる動作概念との対応を明確にし、格フレームを用いて階層的に表現する。そして、この格フレーム階層をもとに構成した状態遷移モデルを用いて、時々刻々と変化する数値データ列から身体各部の動作を推定し、格フレームの形式で表現する。最後にこれらを統合し、身体全体の動作を表現する格フレームを生成する。

〈4・1〉格フレーム表現 本手法では、画像中の人物の動作を表現するための形式として、自然言語文の格構造に基づいた格フレームを用いる。ここで格とは、自然言語文

図 6 動作概念の階層

Fig. 6. Hierarchical concepts of motions.

において動詞や形容詞などの述語に係る語句の意味的役割を表すものであり、Fillmore の格文法⁽¹²⁾によれば、動作の対象を表す対象格、動作が行われる場所を表す場所格、動作の起点を表す源泉格など 8 つの種類に分類されている。格フレームは、このような格の種類をスロット名とし、動詞や名詞など文の構成要素を値とするフレーム表現であり、次のような形式である。

[PRED: aruku, AG: man_a, GO-LOC: by(door),
SO-LOC: front(table)]

ここで PRED は述語を表し、AG はその動作主格を、GO は述語に対する目標格を、そして SO は源泉格を表している。さらに、GO-LOC などのように格ラベルの後に値の種類を限定する意味カテゴリ（ここでは場所）を指定することがある。

〈4・2〉格フレームによる動作概念の階層化 まず格フレームを用いて階層的に表現した動作概念の一般形を図 6 に示す。ここで [PRED:verb1, LOC:loc1] という形式の親フレームから、2 つの子フレームが派生している。1 つは、動詞 verb1 に対し新たな動作特徴 f_1 を得てより具体的な動詞 verb2 に詳細化したものであり、それぞれの動詞本来の意味に基づいて様々な特徴や付随する格が考えられる。例えば、動詞 idou_suru（移動する）を親フレームとすると、これに人物位置 loc の移動速度が低速であることを表す動作特徴 move_slow(loc) が加わることで、より具体化した表現の動詞である aruku（歩く）に置き換えることができる。

もう 1 つは場所格をより限定することで詳細化したものであり、この場合子フレーム中の loc2 は、親フレーム中の loc1 よりもカバーする範囲が限定されており、意味的には loc1 ⊃ loc2 のような包含関係となっている。例えば、何らかの動作が行われた場所を詳細化する場合、単に in(room)（部屋の中で）という表現を、より具体的な by(table)（机のそばで）などのような表現に置き換えることが考えられる。

ここで動詞の詳細化に用いる動作特徴は、画像から得られた数値データ列、および与えられた環境知識から抽出可能なものであるが、自然言語処理の分野においては動詞や名詞などの語句の意味を構成する成分として、意味素性とよばれている。人物の動作概念に関連する意味素性の一覧を表 1 に示す。

基本的にはこれらの意味素性が抽出されれば、階層にしたがってより詳細な動詞を選択することになるが、意味素性に

表 1 意味素性
Table 1. Semantic primitives.

意味素性	評価基準 (x)	説明
動きに関する意味素性		
move(loc)	$ \Delta horiz(loc) $	動く
stop(loc)	$ \Delta horiz(loc) $	静止する
move_slow(loc)	$ \Delta horiz(loc) $	ゆっくり動く
move_fast(loc)	$ \Delta horiz(loc) $	速く動く
move_near(ag,obj)	$ \Delta loc_{ag} - loc_{obj} $	近づく
move_apart(ag,obj)	$ \Delta loc_{ag} - loc_{obj} $	離れる
...		
状態に関する意味素性		
high(loc)	vert(loc)	高い
low(loc)	vert(loc)	低い
face(ag, obj)	dir _{ag} · dir _{ag-obj}	顔の向き
near(ag,obj)	$ loc_{ag} - loc_{obj} $	近い
...		
属性に関する意味素性		
prop_operable	-	操作できるもの
prop_readable	-	読めるもの
prop_loadable	-	のせられるもの
...		

よっては単純に真偽を決められない場合も多い。例えば人物 *ag* とある物体 *obj* との接近を表す意味素性 *move_near(ag, obj)* は、単位時間内にその物体との相対距離がどの程度減少したかを量的に評価して、競合する他の子フレームの意味素性と比較し、いずれを選択するか決めなければならない。

そこで、表に示した評価基準 *x* に対して (2) 式のような $[0, 1]$ の値域を持つ Logistic 関数を適用する。Logistic 関数は S 字型の特性を持ち、*x* の値が小さく特徴が現れていない領域では 0 に近い値を示し、*x* がある特定の値を越えて特徴をより顕著に現す領域では、急激に 1 に近い値を示す。

$$f(x) = \frac{1}{1 - Ae^{-Bx}} \dots\dots\dots (2)$$

ここで *A, B* は、関数値の立ち上がりの位置と傾きに係する定数であり、効果的な特徴選択を行うために意味素性ごとに経験的に決定する必要がある。これにより、競合する子フレームのうち、意味素性の評価値が最も大きいものを選択する。また子フレームが一つしかない場合は、値が閾値以上であれば選択する。*move_near* のように複数の物体が対象となりうる場合には、すべての候補について (2) による評価を行った上で、最も値の高いものを選択する。

〈4・3〉 身体各部の格フレームの生成 次に、各々の部位の動きや状態に対する格フレームを生成する方法について述べる。

図 7 は、身体的位置・姿勢の動作概念を格フレームにより表現したものである。まず、最上位には、単にシーン中に人物が存在していることを表す *iru* (いる) という表現があるが、これに移動を表す意味素性 *move* が付加されることで *idou_suru* (移動する) という動詞に置き換わる。ほかにも *seishi_suru* (静止する)、*tasu* (立つ)、*suwaru* (座る) などがある。また、*in(room)* (部屋の中) のような場所格に対しては *by(door)* (ドアのそば) や *front(table)* (机の

図 7 身体動作の階層的な動作概念
Fig. 7. Hierarchical concepts about body motions.

図 8 全身の状態遷移図
Fig. 8. State transition model of body motions.

前) のように、より詳しい表現が可能となっている。
しかしながら、時々刻々と変化する人物の動作を把握するためには、画像中の人物の動作や状態を反映する動的なモデルが必要である。図 7 の格フレーム表現の中には、*aruku* (歩く) のような継続的な動作を表す動詞と、*tatsu* (立つ) のような瞬間的な動作を表す動詞の双方が含まれているが、それぞれを状態遷移モデルにおける個々の状態と、状態間の遷移とに対応させると、図 8 に示すような階層性を持った状態遷移モデルとして表現することが可能である。ここで図中の実線は、格フレーム階層の動詞の詳細化に関わる意味素性に対応しており、また破線の矢印は状態遷移を伴う動詞に対応したものである。

この状態遷移モデルを用いて格フレーム表現を生成するアルゴリズムを以下に示す。

- (1) まず、現在の状態を *s* とする。新たに得られた人物の位置・姿勢に対し、最上位の状態から順に実線に沿って意味素性を評価し、下位の状態への遷移、すなわち状態の詳細化を行う。
- (2) これにより確定した状態 *s'* が現在の状態 *s* と同じであれば、格フレーム表現は生成せず (1) に戻る。
- (3) そうでなければ状態遷移が発生することとなり、この経路 *s* → *s'* において破線で示すような状態遷移が

図 10 手の階層的な動作概念
Fig. 10. Hierarchical concepts about hand motions.

図 9 頭部の階層的な動作概念
Fig. 9. Hierarchical concepts about head motions.

あれば、対応する格フレーム表現を生成する。

- (4) 新たな状態 s' に固有の格フレーム表現を生成し、(1)に戻る。

なお、先述の瞬間動作動詞については、その結果生じる状態を明示的に表現する必要がある場合がある。例えば“立つ”や“座る”に対しては、“立っている”、“座っている”が相当する。このような表現を、述語の相変化として明示的に扱い、tatsu(ASP:resultant) のように表す。

次に、頭部の動きに関する動作概念を図 9 に示す。頭部の向きは、身体の中でも人物の注目対象を示す部位として重要であると考えられる。まず、頭部が動いている (move) 場合には muku (向く) という表現になり、このとき目標格が方向を表す表現に詳細化されて“上を向く”というような表現となる。また視線方向に何らかの対象物 (obj) がある場合には miru (見る) という表現になり、さらに obj が prop_readable という属性を持っていれば、本のように読むことのできるものとして、yomu (読む) という表現に置き換わる。

最後に手の動作概念を図 10 に示す。人物の動作には手を使用するもの、とくに何らかの対象物に対して作用するものが多く、そのため自然言語の表現も対象物の種類や属性、動作中の位置変化などによって様々である。

本手法では手に関する動作については、〈3・2〉で求めた“取る”、“置く”に対応する toru, oku があり、oku については、対象物 obj_1 を置いた場所が別の物体 obj_2 の上であることを表す on(obj_1 , obj_2) を満し、かつ obj_2 が机のよ

うにもの載せることができる属性 (prop_loadable) をもつ場合には、nosuru (載せる) という表現に置き換わる。

また、手の位置と環境中の物体との接触を検出して得られる sawaru (触る) については、対象物の属性によって siyuu_suru (使用する)、またパソコンのような対象に対しては sousa_suru (操作する) のように詳細化がなされる。

なお、頭部と手についても、図 8 と同様の状態遷移により格フレームを生成する。

〈4・4〉 身体各部の格フレームの統合 以上のようにして身体各部の動作に関する格フレームが生成されるので、これを身体全体の動きを表す 1 つの格フレームに統合する。ただし、そのままでは各部位の表現の詳細さに片寄りがあり、自然言語の表現としては必ずしも適切ではないため、統合の際には動作概念の種類に応じて以下の 3 つの統合パターンを使い分ける。

- (1) より詳しい動作概念に代表させる。
例) 座っている + 本を読む ⇒ 本を読む
- (2) 並列に結合する。
例) 時計を見る + 移動する
⇒ 時計を見ながら移動する
- (3) 動作概念を結合し、新たな動作概念を生成する。
例) ノート PC を持つ + 移動する
⇒ ノート PC を運ぶ

本手法で対象としている動作概念の組合せを考えた場合、ほとんどは (1) のようにより詳しい動作概念に代表させることで、自然な表現を生成することが可能である。そこで、通常は (1) のパターンの結合規則を適用し、特定の動作概念の組み合わせに対して (2) または (3) を適用することとする。表 2 に、それぞれのパターンに対する代表的な結合規則を示す。

5. 自然言語テキストへの変換

ここでは、生成された格フレームを自然言語のテキストへ変換する方法⁽¹³⁾ について述べる。格フレームは、動詞とその他の格との関係を表現したものであるが、これを日本語のテキストに変換する際には、“が格”や“を格”を補う必要がある。このような後置詞パターンは動詞によって

表 2 身体各部の格フレームの統合規則

Table 2. Rules for integrating case frames of body parts.
パターン (1) による統合
body:[PRED:tatsu(ASP:resultant), AG:ag, LOC:loc, ...] + hand:[PRED:toru, OBJ:obj] ⇒ [PRED:toru, AG:ag, OBJ:obj, LOC:loc, ...]
パターン (2) による統合
body:[PRED:idou.suru, AG:ag, ...] + head:[PRED:miru, OBJ:obj] ⇒ [PRED:idou.suru, AG:ag, ...SUB:[PRED:miru, OBJ:obj]]
パターン (3) による統合
body:[PRED:idou.suru, AG:ag, ...] + hand:[PRED:motsu, OBJ:obj] ⇒ [PRED:hakobu, AG:ag, OBJ:obj, ...]

決まっており、分類されている⁽¹⁴⁾。
例えば、“歩く”や“立つ”、“座る”などは、動作主格のみが常に伴って用いられる（必須格）ため、「人物 A が歩く。」のように“が”を後置詞として補う格構造パターンである。また、“読む”や“操作する”などの動詞は、動作主格と対象格が必須格であり、「人物 A が本を読む。」のように“が”と“を”を補う格構造パターンである。このように動詞ごとの格構造パターンに基づいて後置詞を補い、述語節を生成することができる。

また、「人物 A がイスに座る。」のような場合、“イスに”は“座る”にとって必ずしも必要でないため、このような格を自由格と呼ぶ。自由格については個別に後置詞を補う。
次に、格フレームを表 3 に示す構文規則に従って自然言語文に変換する。表 3 の t_0 は動詞であり、 $K_j (j = 1, \dots)$ は動詞を修飾する格のグラベル、 $t_j (j = 1, \dots)$ は動詞を修飾する名詞句を表している。
以上のようにして格フレームから自然言語文に変換した例を以下に示す。

なお、本手法では、動画像からテキスト生成する状況を考えて違和感のないように、動詞の時制としては原則として過去形を用いている。ただし、〈4・3〉で述べた瞬間動作動詞の結果状態（結果相）については、“ている”形としている。

6. 実験および検討

本手法の有効性を確認するために実験を行なった。実験には、本研究室のマシナールームを使用し、室内での動作を想定して動詞を 30 種類、物体を 9 種類とした。画像処

表 3 構文規則

Table 3. Syntactic rules.
(1) <文 [PRED : t_0 , K_1 : t_1 , K_2 : [PRED : t_{02} , K : t]] > ::=< 述語節 [PRED : t_0 , K : t] >< 後置詞 (K_2) > <文 [PRED : t_0 , K_1 : t_1] >
(2) <文 [PRED : t_0 , K : t] > ::=< 述語節 [PRED : t_0 , K : t] >< 終止記号 >
(3) <述語節 [PRED : t_0 , K_1 : t_1 , K_2 : t_2] > ::=< 名詞句 (t_2) >< 後置詞 (K_2) > <述語節 [PRED : t_0 , K_1 : t_1] >
(4) <述語節 [PRED : t_0 , K_1 : t_1] > ::=< 名詞句 (t_1) >< 後置詞 (K_1) >< 動詞 (t_0) >
(5) <名詞句 (t_1) >::=< 名詞 (t_1) >

理とテキスト生成には、PentiumII 266 × 2 の PC 上に C++ と Prolog で構築したシステムを使用した。処理能力は 160 × 120 の画像に対して 2 frame/sec である。ただしこの処理速度では、素早い動作については十分な認識を行うことができなかったため、それぞれのシーンを録画し、1/3スロー再生で行ない、実質処理速度を 6 frame/sec としている。

〈6・1〉 実験結果 10 種類のシーンを設定し、それぞれ 2 回づつ撮影して実験を行った。このシーン 1 つから人物の行動をテキスト表現した結果を図 11 に示す。また、各シーンにおいて生成された動詞が、入力画像と比べて適切であった割合を表 4 に示す。“操作する”の成功率が低いのは、操作対象を置いた時点の位置と、操作時に推定される手の位置のズレが大きくなってしまったためである。この他、たとえば“移動する”については、動作認識としてはほぼ 100% であったが、テキスト生成の過程で別の表現に置き換わる場合が多かったため省略した。また“近づく”、“離れる”など主観的な解釈が必要なものは定量的な評価が難しいため、ここでは評価対象としなかった。

実験結果を見ると、望ましいテキスト表現が得られなかった部分については、画像からの動作推定の不安定さに起因するものが多かった。特に単眼カメラを用いていることによる奥行き方向の位置のズレが影響した。これについては、ステレオ画像など最新の技術を導入すること、高速なハードウェアを利用することなどにより改善が可能であると考えている。

〈6・2〉 検 討 以上の実験結果に基づき、本手法の有効性・新規性について検討を行った。

〈6・2・1〉 物体への作用 一般に人物の動作のうち物体への作用を伴うものとしては、物理的な移動、変形、性質や数量の変化などがある。本研究では、画像から比較的容易に推定できるものとして、物理的な移動のうちでも“置く”、“取る”を対象としているが、このような画像からの情報のみならず、移動の起点・終点など位置情報や、領域知識として与えられる対象物の属性なども含めて評価することで、“載せる”、“操作する”など自然言語の持つ多様な表現を生成する方法を明らかにしている。

図 11 入力画像と生成されたテキスト
Fig. 11. Input images and generated text.

表 4 動詞の生成結果

Table 4. Results of generated verbs.

動詞	成功率 (%)	動詞	成功率 (%)
立つ	93.8	置く	79.4
座る	100.0	載せる	69.9
取る	87.5	操作する	20.0

また変化領域画像による“置く”、“取る”の推定は、シーンに想定される物体が既知である場合には、実験結果で示したように 8～9 割程度の認識率を得ることができた。誤認識となったケースの多くは物体の識別に失敗したものであり今後の改善の余地があるが、本手法の基本的な考え方の有効性は示すことができた。

〈6・2・2〉身体各部の動作概念 本研究では、全身の位置・姿勢と頭部（視線）の向き、手の位置など、身体において機能的な働きを持つ部位を選び、これにより身体全体の動作の推定を試みている。同様に動画からテキストを生成する研究^{(5) (8)}においては、人物の全身あるいは単一の部位により全体の行動の表現を試みているが、人物の動作の

多くは、何らかの対象物との相互作用に基づいており、その対象を特定することにより多様なテキスト表現が可能になる。このため、注視している対象や、手を用いた作用対象を特定することが必要であり、本手法による頭部の位置・姿勢推定、手の位置推定が効果的であると考えられる。

また、頭部や手はそれぞれ独立して動かすことができるため、いずれの動作に注目すべきかはその時々によって変化する。本手法では、各部位から得られた表現を単に並べるだけでなく、詳細な表現を優先したり、複数の部位に関連した動作概念を結合して新たな概念に置き換えるなど、自然言語の動作概念の分析に基づいた統合手法を明らかにしている。

〈6・2・3〉動作概念の階層性 本手法では、自然言語の事象概念について動詞と場所の階層性に注目して格フレームを構成している。動詞については実験結果で示したように、画像から得られる情報や、物体の属性など領域知識を利用して、得られた情報の詳細さに応じた表現を生成することが確認できた。例えば今回の実験では、対象物はすべて机の上に載せる動作を行ったが、このうち置いた物体とその位置の関係が on(obj, table) という意味素性を満たしていることが確認できたのは約 88% で、その場合にのみ“机に載せた”という表現になり、それ以外は単に“置いた”となった。しかし“置いた”という表現でも具体性がやや落ちるものの十分に動作の内容を示しており、本手法で提案する動作概念の階層性が有効に働いた例と考えられる。

一方、場所の階層性については、今回の手法では人物の位置とシーンに配置された机などの物体との間で near（近さ）などの意味素性を評価して決めている。実験結果は図 11 に示した通り全体としてはほぼ行動の流れに沿ったものとなった。このうち“～へ近づく”、“～で立っている”など、人物の微妙な動きに反応してやや冗長な表現も見られるが、本手法では新しい意味素性が抽出されるなど条件に変化があった場合には、逐次テキストとして生成する方針でシステムを構成した。これについてはむしろ、動作表現を時系列上で要約するなどの新たな視点が必要であると考えたためである。

〈6・2・4〉画像の解釈とテキスト生成 これまでも、動画画像からテキストを生成する研究が行われてきたが^{(4) (5)}、これらは監視システムなどにおいて報知内容を人間にわかりやすいようにテキストの形で伝えることを目的としたものである。このため、生成されるテキストが自然言語として豊かな表現を持つことには必ずしも主眼は置かれず、画像の内容が特定の条件を満たした際に表現が一意に定まれば十分目的に沿うものであった。

これに対し本研究では、自然言語の動作概念の階層性に注目し、画像から得られる同じ動作についても、その対象物の種類や属性などの領域知識を補うことで、より自然な表現のテキストを生成することを目的としている。

本手法で用いた動作概念の階層と、それを推定する基準となる意味素性は、現状では室内の人物動作を想定したも

のであり必ずしも一般的なものではないが、具体的な事例に基づいてこのような研究を進めていくことで、画像から得られる動作特徴の抽出法や概念階層の構築法の一般化を図り、将来的にはライブ映像から解説文を生成するなどの応用が可能であると考えている。

7. む す び

本研究では、人物の位置・姿勢や頭部の方向、手の位置といった身体の部位の動きを、自然言語の動作概念に対応付け、人物の行動を適切に表現したテキストを生成する手法を提案した。また、室内環境を対象とした実験を行ない、その有効性を確認した。

謝 辞 本研究の本質的な部分である、画像情報から記号・言語情報への変換について有益なご議論、アドバイスをいただきました大阪大学産業科学研究所の北橋忠宏教授に、また筆者が日頃御指導頂いている本学総合情報センターの田村武志教授に、心より感謝致します。

(平成12年2月29日受付、同12年6月29日再受付)

文 献

- (1) 北橋忠宏：“メディア変換から見た言語と画像および認識・理解の役割”，信学技報，NLC95-39，PRU95-144 (1995)。
- (2) 黄瀬浩一，小島篤博，福永邦雄：“画像・図と言語の統合”，システム/制御/情報，Vol.41，No.12，pp.28-35 (1997)。
- (3) H.-H.Nagel: “A Vision of ‘Vision and Language’ Comprises Action: An Example from Road Traffic,” Artificial Intelligence Review, 8, pp.189-214 (1994)。
- (4) H.Kollnig, H.-H.Nagel and M.Otte: “Association of motion verbs with vehicle movements extracted from dense optical flow fields,” Proc. of 3rd European Conf. on Computer Vision '94, Vol.II, pp.338-347 (1994)。
- (5) M.Thonnat and N.Rota: “Image Understanding for Visual Surveillance Applications,” Proc. of 3rd Int. Workshop on Cooperative Distributed Vision, pp.51-82 (1999)。
- (6) 岡田直之，田町常夫：“自然言語および図形解釈のための単純事象概念の分析および分類”，信学論 (D)，Vol.J56-D，No.9，pp.523-530 (1973)。
- (7) 岡田直之，田町常夫：“動図形の意味解釈とその自然言語記述 — 意味分析 —”，信学論 (D)，Vol.J59-D，No.5，pp.331-338 (1976)。
- (8) 小島篤博，田原典枝，田村武志，福永邦雄：“動画像における人物行動の自然言語による表現”，信学論 (D-II)，Vol.J81-D-II，No.8，pp.1867-1875 (1998)。
- (9) 浅沼克紀，大西正輝，小島篤博，福永邦雄：“色情報と領域追跡情報をを用いた人物の顔と手の領域の抽出”，電学論 (C)，Vol.119-C，No.11，pp.1351-1358 (1999)。
- (10) G.Shafer, “A Mathematical Theory of Evidence,” Princeton University Press (1976)。
- (11) 山中俊夫：“色彩学の基礎”，文化書房博文社 (1997)。
- (12) 田中春美，舟木道雄訳，C.J.Fillmore 著：“格文法の原理”，三省堂 (1975)。
- (13) 高松忍，西田富士夫：“動詞パターンと格構造に基づく英日機械翻訳”，信学論 (D)，Vol.J64-D，No.9，pp.815-822 (1981)。
- (14) 水谷静夫他：“文法と意味 I”，朝倉書店 (1983)。

小 島 篤 博 (正員)

1967年生。1990年3月大阪府立大学工学部電気工学科卒業。1992年3月同大学院博士前期課程修了。同年富士通(株)入社。1996年9月より大阪府立大学総合情報センター助手。画像情報から自然言語情報への変換に関する研究に従事。情報処理学会，人工知能学会，電子情報通信学会各会員。

浅 沼 克 紀 (非会員)

1974年生。1998年3月大阪府立大学工学部情報工学科卒業。2000年3月同大学院博士前期課程修了。在学中はコンピュータビジョンの研究に従事。現在，(株)日立製作所勤務。

福 永 邦 雄 (正員)

1945年生。1967年3月大阪府立大学工学部電気工学科卒業。1969年3月同大学院修士課程修了。同年同大学工学部電気工学科助手。現在同大学工学研究科情報工学分野教授。コンピュータビジョン，グラフ理論とその応用などの研究に従事。情報処理学会，人工知能学会，システム制御情報学会，IEEE 各会員。工学博士。