特定空間内の知識を利用した画像処理結果の言語化

能見　麻未† 　　小林　一郎‡

†お茶の水女子大学　理学部情報科学科　小林研究室　〒112-8610 東京都文京区大塚 2-1-1
‡お茶の水女子大学 大学院 人間文化創成科学研究科 〒112-8610 東京都文京区大塚 2-1-1
E-mail:　†‡{mami_n, koba }@koba.is.ocha.ac.jp,

あらまし　近年，デジタルカメラや Web カメラの普及により，動画像の使用が容易になってきた．しかし，撮影された多量の動画像の中から特定の一部分だけを探し出すことは困難であり，現状では，撮影された内容を人が確認しながら探すことしかできない．

このことから，本研究では取得された動画像に対して画像処理を施し，特定空間内での人の動きとその空間に存在する物体とのインタラクションを観察することにより，人の行為を言葉で説明する手法を提案する．それにより，動画像の中から人の行為を言葉で検索するシステムを開発することを目的とする．

キーワード 画像理解，言語による説明，特定空間における知識，Activity Theory

Verbalizing the Result of Image Understanding with the Knowledge of a Particular Space

Mami NOMI† and Ichiro KOBAYASHI‡

† Department of Information Sciences, Faculty of Science, Ochanomizu University
Otsuka 2-1-1, Bunkyou-ku, Tokyo, 112-8610 Japan
‡ Ochanomizu University Graduate School of Humanities and Sciences
Otsuka 2-1-1, Bunkyou-ku, Tokyo, 112-8610 Japan
E-mail:　†‡{mami_n, koba}@koba.is.ocha.ac.jp

Abstract　Recently as digital cameras and web cameras have been commonly used in our everyday lives, we can easily obtain quite a few movies. However, it is difficult to find a particular part of the obtained movies. From this, in this paper we apply image processing for the obtained movies and then propose a method to explain human's behavior in a particular space by observing how a human beings interacts the objects in the space. By this, we aim to develop a system that enables us to retrieve a particular human behavior by words.

Keyword　Image Understanding, Explaining by words, Knowledge of a Particular Space, Activity Theory

1. はじめに

近年，デジタルカメラや Web カメラの普及により，動画像の使用が容易になってきた．また，インターネットの普及により，防犯や子供，ペットの留守番を遠隔から見ることも可能となっている．しかし，このような長時間撮影を続けている動画像の中から，特定の一部分だけを探し出すことは未だに困難である．

現状では，撮影された内容を人が確認しながら探し出す作業が必要となっている．確認するためには，動画像撮影と同等の時間と手間を要し，動画像が多量になればなるほど困難を極めてくる．そのため，多量の動画像の中から，特定の部分だけを探し出す手法が必要とされる．このことから，本研究では取得された動画像に対し，画像処理技術を施し，動画像内での出来事を言葉で説明する手法を提案する．特に，本研究では，特定空間内での動画像を取得し，人の動きを動画像処理により認識し，人とその空間内に存在する物体とのインタラクションを理解することにより，人の行動に対する言語化を行う．それにより，身近である言葉によって，容易に動画像の欲する特定部分を探し出す手法の提案を目的とする．

2. 関連研究

先行研究として，同様に空間内の人物の行動を言語化する研究が行われている[1]. 檜山ら[1]は，人を示す画像データの特徴として肌色に着目し，肌色抽出を行うことにより人の行動を捉えている．顔や手の肌色部分を特徴データとして扱い，それらと物体とのインタラクションに基づいた言語化を行っている．しかし，肌色部分の抽出だけでは季節や光源の変化により，肌

色部分の露出度や色合いが変化すると共に，後方からの画像からでは特徴データを得ることができないなどの問題点が挙げられる．また，物体とのインタラクションを捉えるために，存在する物体の特定空間内の定義法もあらかじめ座標値を組み込むなど汎用性に乏しいものであった．

このため，本研究では，データの抽出方法と特定空間内における物体の定義方法の改善を行っている．

関連研究においては，人と物との関連を処理順序でまとめた environmental map を用いた画像理解を行う研究[2]や画像からの言語生成ではなく，物にセンサを取り付け，センサから得たデータによって起こる出来事を言語化する研究が行われている[3]．また，言語化を行うだけでなく，言語情報と時間情報を結びつけることにより，有効に利用できるコンテンツを生成し，それを利用した実際の Web アプリケーションも研究開発している．

一方，本研究では固定された Web カメラ 1 台から得られた 2 次元の動画像を用い，人の行動を説明する，言語生成を行う．

3. 言語化システムの構築

図 1　元画像

本研究では，動画像ファイルの初期画像を「元画像（図 1)」として扱う．元画像からその画像を示す空間内に存在する物体を定義する知識を作成し，人の振る舞いは，元画像と入力画像の背景差分を用いて捉え，定義された物体の知識を用いて画像理解結果を言語化するシステムを構築する．

3.1 空間知識の作成

多様な背景（元画像）に柔軟に対応するため，空間内の物体に対する定義法をマウスによる座標指定とする．元画像が表示されているウインドウ上で空間に存在する物体の四隅をクリックすることで（図 2)，対象となる物体の画像内における座標値を取得する．この取得した座標値に定義物体名を付けてファイルへ保存する（図 3)．

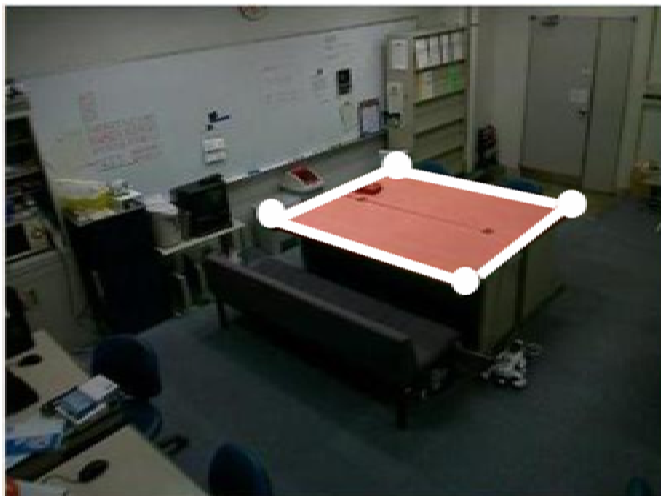

図 2 マウスによる物体定義

```
＊定義設定＊
 マウスによって4点座標を指定する。
 'z'で定義名入力。

 4点座標指定してください。
終了する場合は'a'
1 : 234 202
2 : 160 170
3 : 278 128
4 : 343 179
zを押し、定義名を入力してください。
保存する定義名を入力してください>机
(1)定義名：机
 [定義範囲]
 x座標：160～343
 y座標：128～202
        ファイルに書き込みました。
4点座標指定してください。
終了する場合は'a'
```

図 3　定義座標値の表示画面

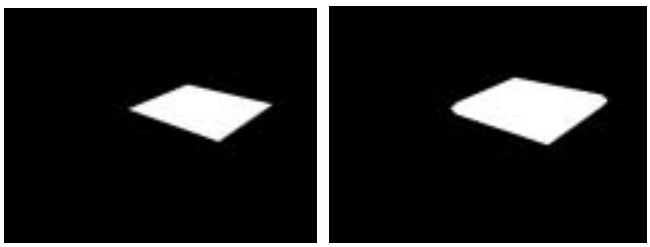

図 4 定義物体の二値化画像
（左図：膨張処理無し　右図：膨張処理有り）

このように空間内に存在する物体の座標値を取得し，次に，システム内で使用する定義域を設定する．

取得された物体の座標値を用いて，定義物体それぞれの二値化画像を作成する．この二値化画像は，特定空間内に存在するどの定義物体に関連した動作を行っているのかを見つけ出すために使用する．また，物体周辺での行動も考えられるため，二値化画像には膨張処理を施し，物体より多少大きく領域の作成を行う（図４）．白の領域が定義物体内，黒の領域が定義物体外に相当し，以下，白の領域である定義物体内を，指定した物体の「定義域」と呼ぶ．

3.2 動画像からの特徴データ抽出

画像認識には，Intel 社が公開している画像処理ライブラリである OpenCV[4]を用いる．提供される各種画像処理法の中の，背景差分，輪郭検出を用いて人の振る舞いを認識する．

3.2.1 背景差分法

背景差分法とは，元画像と入力画像の差分をとり，変化のある画素を抽出する手法である．本研究では，さまざまな室内の空間に適用可能な手法を開発する目的の下，システム開発を行っているため，状況の変化に柔軟に対応できる背景差分法を用いた．誰もいない状況の画像（元画像）と人が存在する画像（入力画像）を照らし合わせることで，人が動いている部分など，変化のある部分だけの画像を抽出することができる．

しかし，背景差分法だけでは，光の変化や影まで差分と捉えてしまうという問題点がある．光や影によって変化する範囲は大幅なものではなく，多少変化したものが差分として捉えられているため，実際には不要な部分が差分として抽出されてしまう．このため，各画素の RGB 値に注目し，元画像:hと入力画像:f，両方の(x, y)座標について RGB 値の比較を行う．

$$R(x,y) = | f_R(x,y) - h_R(x,y) |$$
$$G(x,y) = | f_G(x,y) - h_G(x,y) |$$
$$B(x,y) = | f_B(x,y) - h_B(x,y) |$$

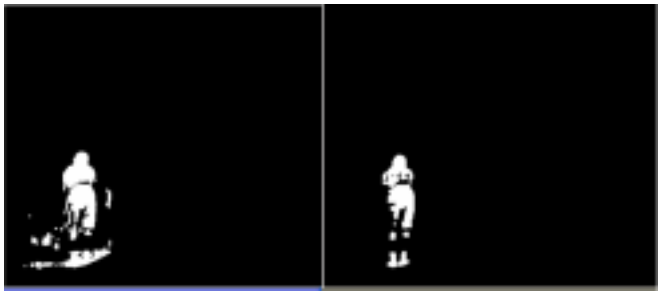

図 5 左図：背景差分法　右図：RGB 背景差分法

元画像と入力画像の R,G,B の値の差が全てある閾値以下であれば，その部分を光や影とみなし，差分から排除する．現在，閾値として「30」と設定している．背景差分法によって抽出された画素値に対してさらに RGB 背景差分法を適用することによって，従来取り除くことができなかった光や影の変化を排除し，より正確な画像を抽出することが可能となった（図 5）．

3.2.2 特徴データ抽出

背景差分法を用いて得られた画像から特徴データの抽出を行う．本研究では，差分で得られた画素を一つの集まりとして捉える事を可能とする輪郭検出を用いる．この捉えた領域の位置情報から重心を計算し，特徴データとして扱う．輪郭線で囲まれる領域にある点の重心を座標値の相加平均として計算し，動画像から得られるすべての画像に対してデータの抽出を行う．

このデータを用いて，画像理解に基づく，言語化を行う．

3.3 事前知識を利用した言語化

画像理解結果の言語化は，予めシステムに付与される，物と人の行動のインタラクションに関する事前知識を用いることにより行う．この事前知識は，ロシアの心理学者 Vygotsky によって提案された Activity Theory[5]を用いる．Activity Theory では，人と物とのインタラクションの分析を，人工物(Artifact)が人の行動の媒介になるという視点から行い，その関係を明らかにする．

このことから本研究では，Activity Theory を参考にし，人の行動に関する分析を行った．例えば,「人が部屋へ入る」という行動に対しての動作は

(1) ドアを開ける
(2) 人が入る
(3) ドアを閉める

この 3 つの動作が行われる．そして，部屋へ入るという行動には,“ 人が入室するため ”という目的と“ ドア ”という空間内に存在する物体が関係する．

表 1 本研究で用いる事前知識（一部）

Scenario	Type	Artifact	Purpose
部屋に入る・出る ・ドアを開ける ・人が入る ・ドアを閉める	Action Operation	ドア	入退室するため
物を冷やす ・開ける ・入れる ・閉める	Action Operation	冷蔵庫	物を冷やし，保管するため
本を取りだす ・本を取る	Action Operation	本棚	本を整理し，保管するため

このように，空間内で行われる人の行動に関して，動作，目的，対象物体の分析を行った．本研究で用いる事前知識の一部を表 1 に示す．

この作成した事前知識に基づき，画像理解結果とそれを説明する適切な言葉を選択する知識をシステム内に構築する．

3.4 言語化システム

画像処理を施し得た特徴データの重心座標が，連続した一定時間，特定空間内で定義した物体領域の中に存在する条件を満たした際に言語化を行う．連続した一定時間とは，時間の開きが存在すると別の動作となることがあるため，10 フレーム以上の開きのない連続した 30 フレーム間（つまり約 1 秒間）として設定する．

定義域内に重心が入ってくると，それぞれの定義物体に付与されているカウンタが増加してゆく．この定義域と重心の関係は，物体定義時に領域を保存しておいた画像と照らし合わせることによってどの物体の定義域内に重心が存在するかを見つけ出す方法をとる．定義域と合致し，増加していったカウンタが 30 の倍数に達したとき，言語化へ結びつける．この際，時間を考慮に入れる．上述の理由により，時間差が存在したときに，カウンタは再び初期化される．

また，同じ動作が続く場合，通常，同一人物が行っているとみなすことができるため，一度言語表示をしてしまえば，約 1 秒ごとに同じ言語表示を繰り返し行うことは冗長となる．そのため，同じ言語表示が続くことを避けるために，一定時間を開けて言語化を行う．現在，120 フレーム，約 4 秒間の時間を開けている．

ここで，定義域と合致したカウンタだけで言語化を行うと，“ドア”に関する言語では，「開ける」と「閉める」の両方が同時に言語化されてしまうという問題点が考えられる．「閉める」という動作には，「開ける」という前提条件が存在するため，そのような前提条件を考慮に入れて言語化を行う．本研究では，前提条件が必要な言葉を書き出す際には，前提条件と合致する言語表現が既出の場合にのみ出力するとしている．

出力される言語表現には，Activity Theory に基づき，基本的には「誰が何のために何をどうする」という形で表示する．ただし，「何を」という対象物体によって助詞の表現方法に違いが出てくるため，上記の表現形式に当てはまらない場合がある．この違いは，体のどの部分を使って動作を行ったかによって，大きく分類することができる．手を動作に使う場合は，「何－を－どうする」というように言語化による表現形式を統一でき，使わない場合は「何－に・へ・から－どうする」と助詞の変化によって多様な表現ができる．よって，動作に使う道具として体の一部分の使用，未使用を明記し，本研究では，手を使用した動作との違いにより，人の行動を分類し，手の使用に関する動作を適切な表現選択の基準として言語表現の文章化を行った．手を使用した“開ける”と，手を使用しない“座る”を例にとると，以下のようになる．

・人 が 入退室のため に ドア の 扉 を 開ける
・人 が 休むため に 椅子 に 座る

4. 実験結果

提案手法を用いて，特定空間内での人の行動を画像理解結果から説明する実験を行った．今回，用いた空間における環境は，ドア，机，椅子，冷蔵庫などがある環境であり，カメラを上方に設置し，固定された一方向から録画された動画像を用いた．

図 6 空間内の定義物体

本実験では，人がドアから入室してきて，机を拭くという作業を行った場合の動画像を使用し，空間内に定義した物体は，「ドア」，「机」，「冷蔵庫」の 3 点である．

4.1 リアルタイムでの言語化

動画像を再生しながら言語化をリアルタイムで行った場合の出力結果は図 7 のようになった．

時刻 74 人 が 入退室のため に ドア の 扉 を 開ける
時刻 98 人 が 入退室のため に ドア の 扉 を 閉める
時刻 177 人 が 作業するため に 机 で 作業している
時刻 535 人 が 作業するため に 机 で 作業している
時刻 658 人 が 作業するため に 机 で 作業している

図 7 言語化出力結果

「開ける」，「閉める」という前提条件に基づき選択される表現が同時刻で出力されていないことから，条件が考慮され，出力されていることがわかる．また，同表現が続く場合も出力時刻の開きが大きいことから，

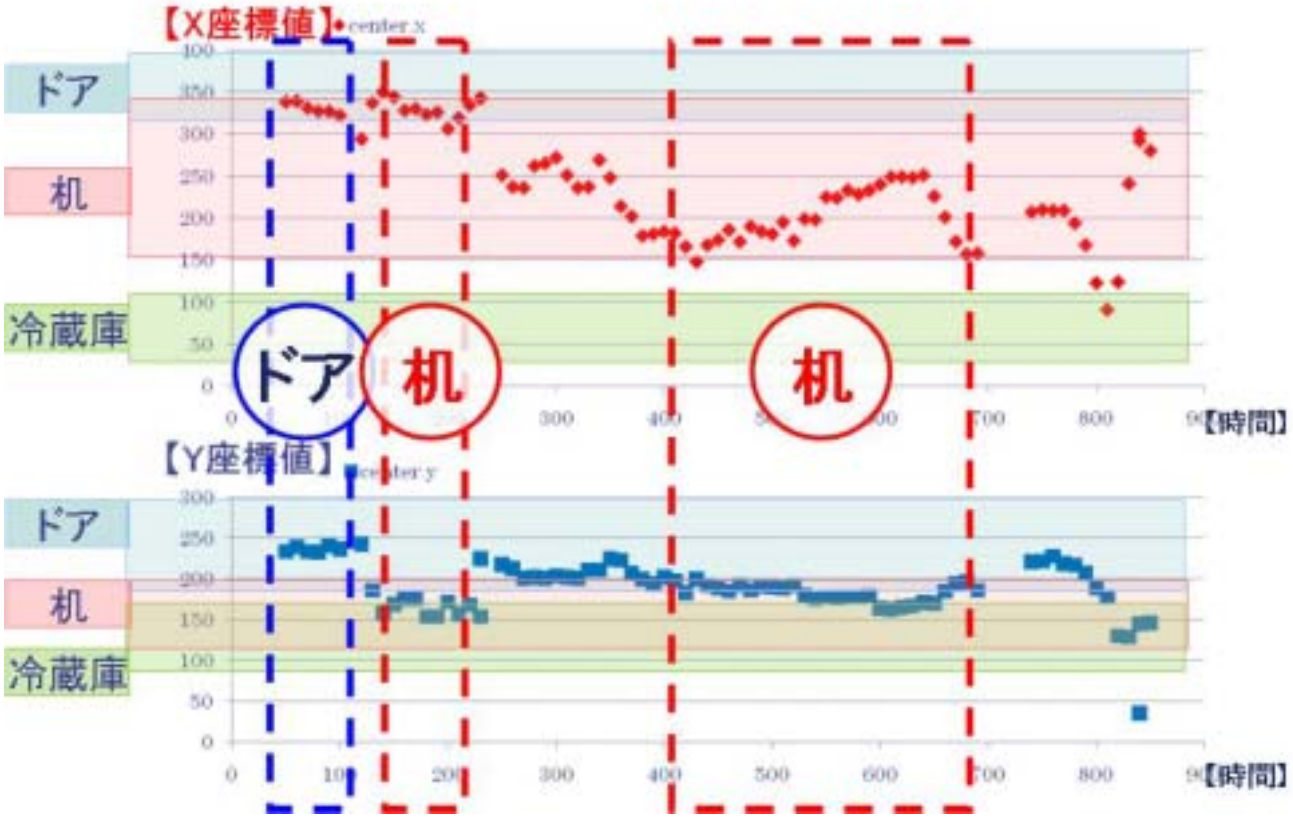

図 8 特徴データのグラフ化

一定時間をおいて言語化を行っていることもわかる．

今回はドアと机にインタラクションする動画像を使用したため，両方に関係する言語化の結果を得ることができた．

4.2 特徴データからのグラフ化

捉えた特徴データの結果を x 座標，y 座標それぞれについてグラフ化し，そのグラフに定義域を重ね合わせたものが図 8 のようになる．

本研究のシステムにおいて，x 座標値，y 座標値が共に，同時刻に同定義域内に存在する際，言語表示が行われることになる．

このグラフから，動画像の前半では，ドア付近での人の動作が観測され，中盤から後半にかけて机付近での人の動作が観測されていることがわかる．

結果として，ドアにインタラクションする人の動作に対する言語化が行われ，その後，机にインタラクションする人の動作に対する言語化が行われた．

よって，リアルタイムで行った言語化が，グラフから判別される人の行動に対する言語化の結果と同じ結果であったと言える．

5. 考察

今回用いた動画像は，ドアと机の 2 つの物体にインタラクションする人の動作を捉えたものであり，複数の画像処理技術を駆使し，動画像内での人の振る舞いを認識し，それに対する言語化を行った．元画像内に定義された物体の領域内に特徴データが合致している場合については，言語化することができた．

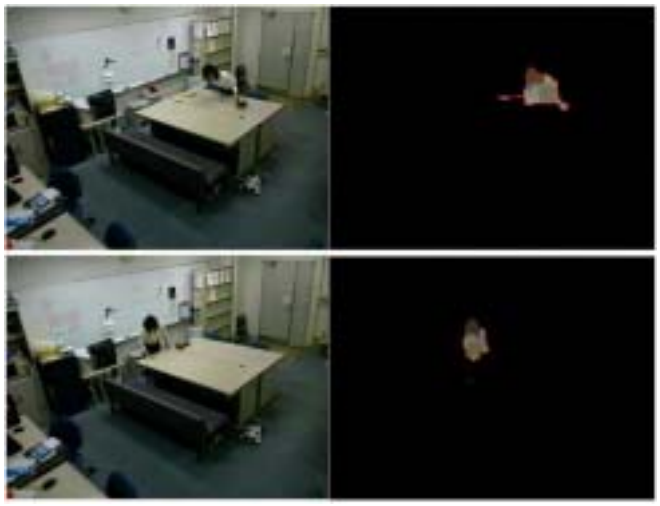

図 9 未認識画像

しかし， 領域内に含まれない定義物体付近で動作を行っている場合については，領域内と認識されないため，定義物とのインタラクションを得ることができない．今回の実験においては，机の周辺を回りながら

机を拭いている画像が認識されない部分に相当する（図 9）．また，グラフによって示される時系列データの変遷を観測することから，より詳細な人の振る舞いを認識できる可能性がある．例えば，観測される時系列データにおいて，ｘ座標は定義域内で変化しているのに対して，ｙ座標はなだらかに定義域の境界付近で推移している場合，人は机の周囲を移動しているのではないかとの推測を得ることができる．

リアルタイムで行った言語化に，グラフから得た結果を付与してゆくことで，人の振る舞いを捉えたより詳細な言語化を行うことが可能になると考える．

6. まとめと今後の課題

本研究では，取得された動画像に対して，画像処理を施し，特定空間内での人の動きとその空間に存在する物体とのインタラクションを観察することにより，人の行為を言葉で説明する手法を提案した．具体的には，空間内の物体定義手法の提案，背景差分画像から動画像内の人の行動を正確に抽出するため，RGB 背景差分法を導入し，背景差分画像認識の精度向上を図った．また，言語化に向けて，Activity Theory による人の行動と物とのインタラクションの分析，それに基づく言語化のための知識作成，それらを用いた画像理解結果の言語化システムの構築を行った．

今後は，さまざまな物体と人がインタラクションを行っている動画像の言語化が行えるように提案手法の汎用性および頑健性を拡張し，言語化された言葉の正確性を向上させると共に，観測された時系列データのグラフ分析に基づいた人の行動の推測などを行い，より詳細な言語表現の付与を行っていきたいと考えている．また，空間内の事前知識の充実化を行い，より豊かな言語表現による説明を目指し，様々な角度からの言葉による動画像検索に応用できるようにするつもりである．

文　献

[1] 檜山　敦子，小林　一郎：“生活空間内における人物行動の画像理解による言語での説明”，情報処理学会全国大会，4P-3，Mar.2007.

[2] Mirai Higuchi，Shigeki Aoki，Atsuhiro Kojima，Kunio Fukunaga：“Scene Recognition Based on Relationship between Human Actions and Objects”，Proceedings of the 17th International Conference on Pattern Recognition (ICPR’04)，2004.

[3] NTT コミュニケーション科学基礎研究所：“S-room 実世界情報の生成とそのリアルタイムコンテンツ化”，NTT 技術ジャーナル，pp.13-18，Jun.2007. http://www.ntt.co.jp/journal/0706/files/jn200706013.pdf

[4] “OpenCV”，http://opencv.jp/

[5] Vygotsky：“Acting With Technology:Activity Theory And Interaction Design(Acting With Technology Series) ”, Oct.2006.

時刻 0

時刻 74 「人が入退室のためにドアの扉を開ける」

時刻 98 「人が入退室のためにドアの扉を閉める」

時刻 177「人が作業するために机で作業している」

時刻 535「人が作業するために机で作業している」

時刻 658「人が作業するために机で作業している」