コンピュータビジョンと
イメージメディア 107－5
（1997. 9. 11）

動画像における人物行動のテキスト表現

田原 典枝† 小島 篤博‡ 田村 武志‡ 福永 邦雄†

† 大阪府立大学工学部
‡ 大阪府立大学総合情報センター

概要 施設の警備および監視，医療機関における看護援助等では，動画像による遠隔監視システムが用いられる．しかし，このようなシステムでは，転送情報の通信容量や監視人員の負担の問題が指摘されている．これらの問題を解決するため，本論文では，動画像中の人物の動きを抽出し，簡潔な説明テキストを生成するとともに，対象物の動き・動作などをテキスト形態で報知する方法を提案している．まず，時系列画像の各フレームに現れる人物頭部の位置・姿勢を推定する．この結果から人物の時系列上の動きを抽出し，いくつかの基本的な動作パターンに分類する．次に，対象人物のおかれている周囲の環境を考慮した推論規則を用いて，目的を持った動作レベルの行動を推論する．さらに，推論した行動や状態を説明するテキストを自然言語処理により生成するとともに，報知する方法を提案している．

Expression of Human Movements on Sequential Images in Text Form

Norie Tahara† Atsuhiro Kojima‡ Takeshi Tamura‡ Kunio Fukunaga†

† College of Engineering, Osaka Prefecture University
‡ Library and Science Information Center, Osaka Prefecture University

Abstract Remote monitoring systems using sequential images are widely used for guarding facilities, nursing in medical facilities and so on. But, it is necessary for those systems to reduce communication capacity and cost of monitoring at all time. Then, this paper proposes a system which expresses human movements on sequential images in test form. Firstly, the system estimates the 3-D position and orientation of the object appeared on each frame of the sequential images. Using the results, the system extracts the primitive human movements and classifies these into model movements. Secondly, the system estimates the objective human movements by combining with the knowledge of the environment using inference rules. Finaly, the reporting sentences are composed.

1 はじめに

近年，コンピュータビジョンの研究において，人物や動物の手足の動きや移動を認識する研究が行なわれており，人物のジェスチャーや顔の表情をヒューマンインターフェイスに応用する研究 [1] や，防犯用警備システムや医療機関・高齢者福祉家庭のための事故防止システム，さらに希少動物の生態観察などを行なうシステムに応用されている．また，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 [2][3] も行なわれている．

従来，人物や動物の行動を監視するシステムでは，監視対象を捉えた動画像を遠隔地のモニタで監視する遠隔監視システムや，特定のセンサーが反応した時に報知するシステムが用いられている．しかし，遠隔監視システムでは，一定時間間隔で得られる複数の画像を全て転送する必要があり，データ量が多く遠隔地への転送に向かない．また，転送され

た動画像を人間が監視する場合, 常に異常を見逃さないように画像を注視する必要があり, 人的負担が大きい.

一方, 私達人間は, 様々な概念を自然言語を用いて表現し, これを声に出したり文字に置き換えることにより物事を他者に伝えることができる. したがって, 自然言語を用いて画像上の動物体の行動や動作を表現することは, 人間に対して動画像上の情報を明確に伝える有効な手段の一つであるといえる.

そこで本研究では, 画像上の対象人物の動きを抽出し, これを簡潔に説明する自然言語テキストを生成することにより, 監視あるいは報知するシステムを提案する.

まず, 固定したカメラから得られる入力動画像上の対象人物の動きを抽出する. ここでは, 対象人物の体全体の移動や手足の動作を抽出しなくても, 対象人物の注意がどこに向いているのかを頭部の位置・姿勢から抽出できると考え, 入力動画像の各フレームから, 画像処理により対象人物頭部の三次元空間における位置・姿勢を推定する. 次に, 得られた位置・姿勢の変化から対象人物の動きをいくつかの基本的な動作に分類する. この基本動作をもとに, 対象人物を含む環境の状態も考慮して, 有目的動作などのより複雑な行動レベルの動作表現を, 推論規則を用いて生成する. さらに, 自然言語処理を用いて, これらの行動表現を簡潔な自然言語による説明テキストに変換する. また, 特定の行動が検知された場合には, この自然言語によるテキスト情報を, 報知システムを通じて遠隔地に通報する (図 1).

以下 2. から 4. で各段階の処理について述べ, 5. で本手法の有効性を確認するために行なった実験について説明する. 最後に, 6. で本研究について考察する.

2 対象人物頭部の位置・姿勢推定

本研究では, 対象人物の動きを抽出するため, 時系列画像の各フレームにおける人物頭部の位置・姿勢を, サンプル画像を用いて推定する. 図 2は本研究で用いた基本サンプル画像である. ところで, 人物頭部の画像上での見え方は, カメラに対する人物の位置や姿勢により様々に変化する. したがって, 基本サンプル画像だけでは人物頭部の多様な位置・

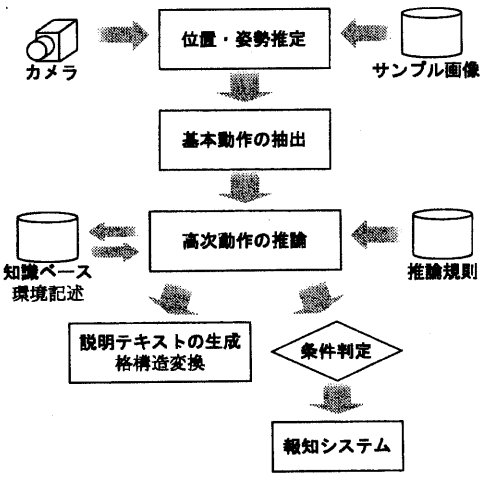


図 1: 処理の流れ

姿勢を推定することができない. そこで, カメラからの位置により大きさの変化する頭部の位置・姿勢を推定するために, 各基本サンプル画像を段階的にリサイズしたものをサンプル画像として用いる. また, これらのサンプル画像をもとに線形変換することにより, 多様な姿勢の画像を生成し, 利用する.

2.1 座標系

図 3に本研究で用いる座標系を示す. $O_c - x_c y_c z_c$ 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 軸とする. $O_m - x_m y_m z_m$ は頭部座標系であり, その原点 O_m を頭部の中心にとることにす

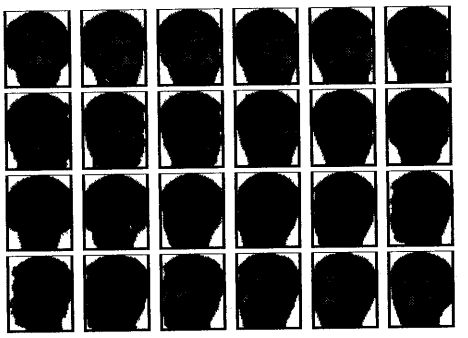


図 2: サンプル画像

る．また，$O-XY$ は画像平面である．

図 3において，対象座標系の各軸 x_m, y_m, z_m に対する対象人物頭部の回転をそれぞれ ϕ, θ, ρ とし，頭部座標系の原点をカメラ座標系における $\boldsymbol{t} \equiv [t_x\, t_y\, t_z]^T$ の位置におく．すなわち，ϕ, θ, ρ の 3 パラメータが対象人物頭部の姿勢を表し，$t_x\, t_y\, t_z$ の 3 パラメータが対象人物頭部の 3 次元位置を表す．そこで，頭部の位置・姿勢パラメータ $\boldsymbol{x}$ を次式で定義する．

$$\boldsymbol{x} \equiv [\phi\, \theta\, \rho\, t_x\, t_y\, t_z]^T \tag{1}$$

このカメラに対する位置・姿勢パラメータを推定すれば，部屋などの対象人物のおかれた 3 次元空間に対する頭部の位置・姿勢を，座標変換により推定することが可能である．

2.2 エッジポテンシャル画像を用いた位置・姿勢推定

各サンプル画像との類似度を比較評価することにより，人物頭部の位置・姿勢を推定する．ここで問題となるのは，濃淡画像は光源の位置や種類により大きく変化することである．そこで，濃淡画像から光源の位置や種類による影響の少ないエッジを抽出する．ここでは，計算が比較的簡便であるにもかかわらず，比較的よい結果の得られる Sobel オペレータを用いてエッジ抽出を行ない，細線化処理を施して [4] 得られる画像をエッジ画像とする．

しかし，入力画像とサンプル画像それぞれから抽出したエッジ画像を比較評価すると，局所最小値に落ち込む可能性が大きい．そこで，価値が局所最小値に落ち込みにくくするため，それぞれの画像の各画素に対し，エッジ点からの距離に応じて，次式に示すようなポテンシャル場を与えておく．

$$U(r) = \frac{k}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{r^2}{2\sigma^2}\right) \tag{2}$$

ただし，

$$r \equiv \sqrt{X^2 + Y^2}$$

はエッジ点からの距離を表す．また，k, σはそれぞれポテンシャルの高さおよび広がりを表す定数である．エッジ画像上の画素$\boldsymbol{X}$におけるポテンシャル $E(\boldsymbol{X})$ はすべてのエッジ点からのポテンシャルの総和であり，次式のように定義する．

$$E(\boldsymbol{X}) = \sum_{i=1}^{n} U\left(\|\boldsymbol{X} - {}^e\boldsymbol{X}_i\|\right) \tag{3}$$

ここで，n は画像中のエッジ点の数を表す．また，${}^e\boldsymbol{X}_i\,(i = 0, 1, \cdots, n)$ は画像中のエッジ点とする．このようにして生成したポテンシャル $E(\boldsymbol{X})$ をエッジポテンシャル画像と呼ぶことにする．図 4にエッジ画像 (a) と，エッジポテンシャル画像 (b)，さらに (b) を 3 次元的に表現した例 (c) を示す．

以上の処理により，各サンプル画像からサンプルエッジポテンシャル画像を生成する．一方，入力動画像の各フレームから，対象人物の頭部をカラー情報を用いて切り出す．次に，切り出した頭部の画像を濃淡画像に変換した画像からエッジ画像を抽出し，サンプルと同様にエッジポテンシャル画像を生成する．このようにして得られた入力エッジポテンシャル画像を，各サンプルエッジポテンシャル画像と比較評価することにより頭部の位置・姿勢を推定する．ここで，各画像をその画素値を要素とするベクトルで表すと，相関の高い画像間の内積は大きくなる [7]．そこで，入力とのエッジポテンシャル画像間の内積が最も大きいサンプル画像の示す位置・姿勢を，対象人物頭部の位置・姿勢とする．

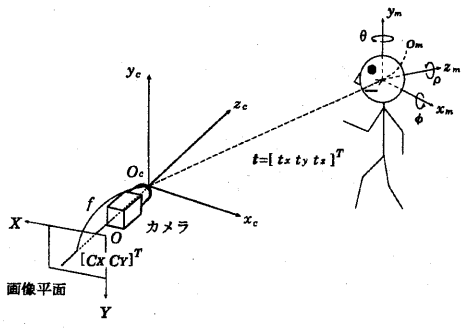
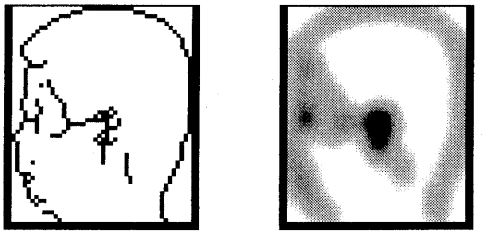


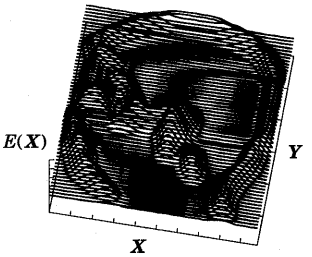
図 3: 座標系

3 対象人物の基本動作の抽出

対象人物を捉えた固定したカメラから得られる時系列動画像の各フレームに対して前述の手法を適用することにより，人物頭部の時系列上の動きを抽出することができる．



(a) (b)



(c)

図 4: エッジポテンシャル画像の例

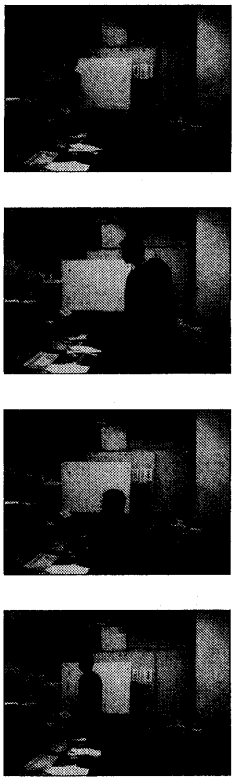


図 5: 動画像の例 1

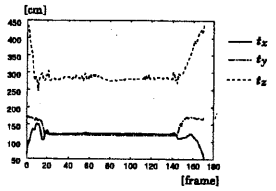
しかし, 画像処理により得られた位置・姿勢推定値は物理量に基づく数値データであるため, 推定誤差や対象人物の動きの不規則な変化による雑音を含み, 対象人物の動きを大局的に抽出するのには適さない. そこで, 得られた位置・姿勢の変化が同じ傾向を維持している区間を基本動作として抽出する.

3.1 動画像における位置・姿勢の変化

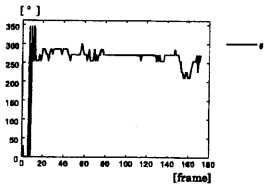
動画像を構成する画像フレーム $I_i\,(i=1,2,3,\cdots)$ 中の対象の位置・姿勢を逐次推定し, これを位置・姿勢パラメータ $\boldsymbol{x}_i$ とする. そして, この位置・姿勢パラメータの変化 $\Delta\boldsymbol{x}_i=(\boldsymbol{x}_{i+1}-\boldsymbol{x}_i)$ が同じ傾向を示す区間を, 基本的な単位動作として抽出し, 幾何学的な特徴に基づいて分類する. ここでは, 対象の種類に依存せず, 単純で一般的なものを動作パターンとして選んだ. これらを基本動作と呼ぶことにする.

図 5 は, 実験対象の人物が, 部屋に入ってきてしばらくワークステーション "eva" を使い, 再び出ていった様子を示している. また, 図 6 はこの時系列画像から推定した位置・姿勢パラメータのうち θ, t_x, t_y, t_z の変化を示している. 対象人物が部屋に入ってきてから "eva" の方へ歩いている間 t_x, t_z が変化し, 同様に "eva" の前から出入り口の方へ歩いている間も, 同様に t_x, t_z が変化している. また, 対象人物が椅子に座る際に t_y が変化し, "eva" を使っている間 θ, t_x, t_y, t_z は, ほぼ一定である.

このように, 時系列画像上における対象人物頭部の位置・姿勢パラメータの変化は, 対象人物の動きを反映している. そこで, 図 6のような位置・姿勢パラメータの変化を, 後述する基準にしたがって分類することによって, 対象人物の基本動作を抽出する.



<位置>



< 姿勢 >

図 6: 位置・姿勢パラメータの変化

3.2 基本動作の抽出

図6に示したようなパラメータの変化は，推定誤差や対象の動作の微妙な変動などによる雑音を含むため，そのままでは対象物本来の動作パターンを正確にとらえられない可能性がある．そこで，基本動作抽出の前処理としてこれらの複数フレーム間のパラメータの変化の平滑化を行なう．

第 i 番目のフレームにおけるパラメータの増分 $\Delta\overline{x}_i$ を次式のように定義する．

$$\begin{aligned}\Delta\overline{x}_i &= (c_k\Delta x_{i+k} + c_{k-1}\Delta x_{i+(k-1)} + \cdots \\ &+ c_0\Delta x_i + c_1\Delta x_{i-1} + \cdots + c_k\Delta x_{i-k})\end{aligned} \tag{4}$$

ただし，

$$c_0 + 2\sum_{j=1}^{k} c_j = 1$$

このように平滑化されたパラメータの増分 $\Delta\overline{x}_i$ の各成分 $\Delta\overline{\phi}, \Delta\overline{p}, \Delta\overline{\theta}, \Delta\overline{t}_x, \Delta\overline{t}_y, \Delta\overline{t}_z$ から，図7に示す条件を用いて基本動作を抽出する．

ここに示す基本動作の条件は，それぞれの動作に特徴的な方向や変化量に対する閾値（X_{TH}）によるものである．なお，いくつかの基本動作の条件に重複を許しているが，より大きな特徴量を有している動作を基本動作としている．観察対象の人物が図5のような動きをした場合を考える．部屋に

go_horizontal(水平方向の移動) :

$|\Delta\overline{t}_y| < y_{TH} \wedge (|\Delta\overline{t}_x| > x_{TH} \vee |\Delta\overline{t}_z| > z_{TH})$

$|\tan\varphi| < 1/\sqrt{3}$ ⇒ 左右の移動

$1/\sqrt{3} < |\tan\varphi| < \sqrt{3}$ ⇒ 斜めの移動

$|\tan\varphi| > \sqrt{3}$ ⇒ 前後の移動

ただし, $\tan\varphi = \Delta\overline{t}_z/\Delta\overline{t}_x$.

go_upward(上への移動) : $\Delta\overline{t}_y > y_{TH}$

go_downward(下への移動) : $\Delta\overline{t}_y < -y_{TH}$

turn(方向転換) : $|\Delta\overline{\theta}| > \theta_{TH}$

stay(静止) : $|\Delta X| < X'_{TH}$,

$(X = \overline{\phi}, \overline{p}, \overline{\theta}, \overline{t}_x, \overline{t}_y, \overline{t}_z)$

ただし，X_{TH} は閾値.

null : 対象が検出されなかった.

図 7: 基本動作抽出のための条件

入ってきてカメラに近付きながら歩いている間，位置パラメータ のうち頭の高さを表すパラメータ t_y の値の変化は小さいが，カメラからの水平方向の距離を表すパラメータ t_x, t_z の値が大きく変化し，基本動作 'go_horizontal' が得られる．その後の動きについても同様に抽出すると，椅子に座る時と椅子から立つ時に 位置パラメータのうち t_y の値が増加し，基本動作 {go_downward, stay, go_upward} のような基本動作が順に得られる．

4 高次動作の推論と説明テキストの生成

3. で述べた処理により，時系列画像中の対象人物の基本動作を抽出することができる．さらに，対象人物の複雑な動作も図7に示した基本動作の組合せにより表現することができると考えられる．しかし，単純に基本動作列をより抽象的な動作表現を用いて表現するだけでは，対象人物の有目的動作を抽出することができない．そこで，対象人物がいる環境，すなわち部屋の出入り口や物の位置，時刻などの情報を考慮する必要がある．ここでは，推論規則を用いて対象人物の多様な有目的動作を推論し，その説明テキストを生成する手法について述べる．以下，基本動作列と環境条件の組合せから得られる対象人物の動作を，高次動作と表現する．

高次動作	推論規則
部屋に入る	null(so-time:t1, go-time:t2) ∧ go_horizontal(ag:a1, so-loc:p1, go-loc:p2, so-time:t2, go-time:t3) ∧ near(loc:p1, obj:door) ⟶ enter(ag:a1, go-loc:room, time:t2)
歩く	go_horizontal(ag:a1, so-loc:p1, go-loc:p2, so-time: t1, go-time:t2) ∧ near(loc:p2,obj:o1) ⟶ walk(ag:a1,go-loc:o1)
座る	go_downward(ag:a1, so-loc:p1, go-loc:p2, so-time:t1, go-time:t2) ∧ front(loc:p2, obj:o1) ∧ height(loc:p2, height:mid) ⟶ sit_down(ag:a1, front-loc:o1, time:t2)

表 1: 推論規則の例

4.1 推論規則による高次動作表現の生成

対象人物の高次な動作を表現するため, その人物のいる部屋の椅子や机, 時計などの位置や時刻などの環境条件と, 得られた基本動作列とを統合する手法について述べる.

記号処理の分野においては, 既存の知識の集合から新たな知識を得るための推論機構について, 数多くの手法が提案されている [8],[9]. これらの推論機構を参考にして, 環境を考慮し, これまでの処理で得られた基本動作表現の集合から, 有目的行動や状態を表す高次動作表現を得るための推論規則を用意する. 推論規則は, 対象物の動作を効果的に表現するのに最も有効な規則を採用している. 表 1 に本研究で用いた推論規則のうち, いくつかを示す.

例として, 基本動作 "stay" が得られ, その時の対象人物の位置が "eva" の前であり, かつ "eva" の方を見ているということが環境情報から得られたとき, 対象人物が "eva" を操作していることを推論する規則を以下に示す.

例:

stay(ag:*ag*, loc:*p*, dir:*d*,
so-time:*t3*, go-time:*t4*) ∧
front(loc:*p*, obj:eva) ∧
dir(loc:*p*, dir:*d*, obj:eva)
→ operate(ag:*ag*, obj:eva,
so-time:*t3*, go-time:*t4*)

4.2 説明テキストの生成

次に, 4.1 の例に示した高次動作表現を, 自然言語の文に変換する. 自然言語処理の分野では, 自然言語文と, その計算機上における意味表現との間の相互変換について, これまで多くの研究がなされてきた. 本研究では, その一手法である格構造変換を用いて, 高次動作表現を自然言語に変換する [10]. ここで "格" とは, 自然言語文において, 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る.

まず, 高次動作の動作名を述語動詞格 (PRED:) に組み入れ, 格構造表現に再構成する. 格構造表現は, 自然言語文の意味構造を, Fillmore の格文法 [11] に基づき, 動詞と動詞に係る語句の格との関係によって表した文の意味表現である. 格構造表現を以下に示す.

$$(K_1 : t_1,\ K_2 : t_2,\ ...,\ K_n : t_n)$$

ここで $t_i\,(i = 1, 2, \cdots, n)$ は文を構成する語句に相当する記号を, $K_i\,(i = 1, 2, \cdots, n)$ は各語句の文中における役割を表す属性すなわち '格' を表す.

例えば,

(PRED:stay, AG:man,
SO-TIME:t3, GO-TIME:t4)

は,「人が時刻 t3 から時刻 t4 までじっとしていた。」という自然言語文と 1 対 1 に対応している.

このような格構造表現に対し，単語辞書，動詞の格構造パターン，および構文規則を適用すると，自然言語文を生成することができる．

5 実験

本手法の有効性を確認するため，一室にビデオカメラを設置し，人物を観察対象として実験を行なった．

ビデオカメラで撮影したカラー画像を，ビデオキャプチャボード経由でPCに毎秒3コマ取り込んだ．各フレームに対して2.で述べた処理を適用することにより，対象人物頭部の位置・姿勢を推定した．

図5のような入力動画像から得られた，位置・姿勢の変化を図6に示す．ただし，ここでは姿勢パラメータのうち，水平方向の回転パラメータ θ のみを示した．

これらのグラフから，3.1で述べた手法により人物の基本動作を抽出した結果を以下に示す．

```
go_horizontal(ag:man,
                so-loc:p1, go-loc:p2,
                    so-time:t1, go-time:t2)
go_downward(ag:man,
                so-loc:p2, go-loc:p3,
                    so-time:t2, go-time:t3)
stay(ag:man, so-loc:p3, go-loc:p4,
                    so-time:t3, go-time:t4)
go_upward(ag:man, so-loc:p4, go-loc:p5,
                    so-time:t4, go-time:t5)
go_horizontal(ag:man,
                so-loc:p5, go-loc:p6,
                    so-time:t5, go-time:t6)
```

さらに，この基本動作列に対し推論規則を適用し，高次動作を推論した結果，次のような結果が得られた．

```
enter(ag:man, go-loc:room, time:t1)
walk(ag:man, go-loc:eva)
sit_down(ag:man, front-loc:eva, time:t3)
operate(ag:man, obj:eva,
                so-time:t3, go-time:t4)
stand_up(ag:man, front-loc:eva, time:t5)
walk(ag:man, go-loc:door)
go_out(ag:man, so-loc:room, time:t6)
```

最後に，格構造変換を行ない，次のような説明テキストが生成された．

“人が時刻 t1 に部屋に入ってきた。”
“人が eva のところまで歩いた。”

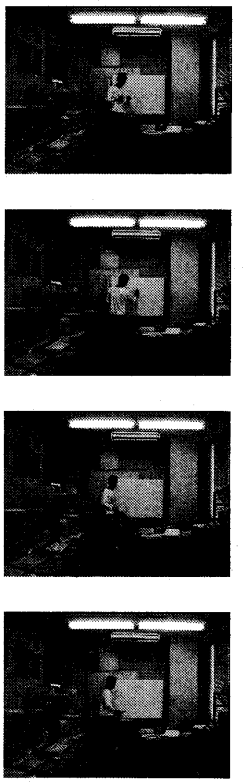


図 8: 動画像の例2

“人が時刻 t3 に eva の前に座った。”
“人が時刻 t3 から時刻 t4 まで
eva を操作していた。”
“人が時刻 t5 に eva の前で立った。”
“人がドアのところまで歩いた。”
“人が時刻 t6 に部屋を出ていった。”

同様に，図8のような入力画像から，次のような説明テキストが生成された．

“人が時刻 t1 から時刻 t10 までホワイトボードに何か書いていた。”

6 考察

図5に示す入力動画像は，対象人物が部屋に入ってきてしばらく “eva” を使った後，出ていく様子を

撮影したものである. 実際のビデオ画像と図 6とを比較すると, 対象人物頭部の位置・姿勢をほぼ追跡できていることがわかる.

また, 位置・姿勢パラメータの変化から抽出した基本動作列と環境情報を用いて高次動作を推論し, 自然言語に変換した結果, 5. に示した結果が得られた. これらを実際のビデオ画像と比較すると, ビデオ画像上の人物の動作を明確に表現する自然言語テキストが生成できていることがわかる.

7 まとめ

本稿では, 動画像中の対象人物の幾何学的な動きを抽出し, 目的をもった行動を推論し, 自然言語による説明テキストを生成する手法を提案した.

より複雑かつ多様な人物行動を抽出するために, 観察対象人物のいる状況に応じた環境情報と基本動作を統合する推論規則を充実することが, 今後の課題として挙げられる.

参考文献

[1] 中村裕一, 西谷正志, 大田友一: "プレゼンテーション映像における話者の行動理解", 信学技報, **NLC95-38, PRU95-143**, p.51-56(1995-10).

[2] D. Koller, N. Heinze and H.-H. Nagel: Algorithmic characterization of vehicle trajectories from image sequences by motion verbs; *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition '91*, pp. 90–95 (1991).

[3] 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).

[4] 高木幹雄, 下田陽久: "画像解析ハンドブック", 東京大学出版会 (1991).

[5] 山根定章, 泉正夫, 福永邦夫: "モデルベースに基づく物体の位置・姿勢推定", 信学論 (D-II),**J79-D-II**,2,pp.165-173(1996-02).

[6] Richard J. Qian, Thomas S. Huang: Object Detection Using Hierarchical MRF and MAP Estimation; *Proceedings Computer Vision and Pattern Recognition '97*, pp. 186-192(1997).

[7] 村瀬洋, シュリーナイヤー: "2次元照合による3次元物体認識–パラメトリック固有空間法–", 信学論 (D-II),**J77-D-II**,11,pp. 2179-2187(1997-11).

[8] 大須賀節雄: "知識情報処理", オーム社 (1986).

[9] 石塚満: "知識の表現と高速推論", 丸善 (1996).

[10] 西田富士夫, 高松忍, 谷忠明: "要求仕様における日本語表現と形式表現間の相互変換", 情報処理学会論文誌, **29**,4, pp.368-377(1975).

[11] 田中春美, 舟木道雄訳, C.J.Fillmore 著: "格文法の原理", 三省堂 (1975).