社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2008－CVIM－163 (14)
2008／5／8

多眼カメラを使った事前学習を要しないアピアランスベース姿勢推定

武田 悟郎† 岡谷 貴之† 出口光一郎†

† 東北大学大学院情報科学研究科
仙台市青葉区荒巻字青葉6－6－01
E-mail: †{takeda,okatani,kodeg}@fractal.is.tohoku.ac.jp

あらまし 画像から物体の姿勢をアピアランスベースで推定する従来の方法は，あらかじめ得た画像と姿勢との学習サンプルを元に推定を行うというものであった．したがって，物体の姿勢の自由度が増すにつれ学習が大変になる（特にサンプルの取得に要する時間）ことや，そもそも学習した物体以外ではうまく推定を行うことができない問題があった．少数のカメラを組み合わせて使うYangらの微分カメラは，これらの問題を解決すべく提案されたものだが，姿勢変化とアピアランスの関係が線形近似できるという仮定に基づいている．両者の関係は本来かなり非線形なものであって，その齟齬からくる手法の限界（推定が行えるシーンの種類等）は明白である．本稿では，微分カメラよりも多量のカメラを使って，線形近似によることなく，直接姿勢変化を推定する方法について述べる．具体的には，基準となるカメラのある時刻の画像と他のカメラの後の時刻の画像とを，射影変換を作用させながら比較することで，物体の姿勢変化を推定する．

Appearance-based Pose Estimation Using Multiple Cameras Without Learning Process

Goro TAKEDA†, Takayuki OKATANI†, and Koichiro DEGUCHI†

† Graduate School of Information Sciences, Tohoku University
6-6-01 Aza Aoba, Aramaki, Aoba-ku, Sendai
E-mail: †{takeda,okatani,kodeg}@fractal.is.tohoku.ac.jp

Abstract Conventional appearance-based methods for estimating the pose of an object from its image use the learning samples of object appearance under known poses that are acquired in advance. These methods suffer from the increase in the time required for acquring the learning samples with the increase of the degrees of freedom of the pose parameters. Furthermore, these methods cannot deal with unknown objects other than the ones that have been learned in advance. The differential camera proposed by Yang et al., which was developed to solve these problems by using a rigid combination of a few cameras, entirely depends on the assumption that the relation between the pose change and the appearance change can be well approximated by a linear function. However, the relation is in essence highly nonlinear, which clearly leads to the limitations of the method. In this paper, we propose a method that uses a larger number of cameras than the differential camera and directly estimates the object pose without the above linearity assumption. To be specific, the method estimates the changes of the object pose by comparing the image captured at a time by a refenrece camera with the images that are captured later by other cameras and are warped by projective transformations.

1. はじめに

画像から3次元物体の姿勢やカメラの相対運動を推定する問題はコンピュータビジョンにとって基本的かつ重要な問題であり，これまでに様々な研究がなされてきた．それらはアピアランスを直接利用する方法 [1] [2] とそうでない方法の2つに分類することができる．

アピアランスを利用しない方法として，モデルベースの方法や対応点を利用した方法がある．

モデルベースの方法は，画像からエッジやシルエットなどの特徴量を抽出し，それをあらかじめ記憶してある形状モデルと照合することにより推定を行う．しかし当然ながらこの方法は，対象となっている物体の形状モデルがない場合には使うことはできない．

対応点を利用する方法は，コーナーなど特徴点を検出し，それらを姿勢変化時の画像間で追跡し，かつ3次元形状を復元することで姿勢変化を推定する．しかし，テクスチャが少ない物体ではそもそも特徴点を追跡することが難しいし，安定して3次元形状を復元することが難しい場合もある．

アピアランスベースの方法は，画像全体の見えを利用した2次元的な照合により推定を行う．パラメトリック固有空間法 [3] などの，今までに提案されているアピアランスベースの推定法の多くは，画像と姿勢との関係をあらかじめ多数学習しておき，その学習データを使って推定を行うという，見えを学習する方法である．しかしこの方法では，物体の姿勢の自由度が増すにつれ学習用の画像を用意することが大変になる．例えば，3自由度分の姿勢を推定するのに1自由度につき10のサンプルを取得したとすると，画像取得に合計 $10^3 = 1000$ 枚必要となり，6自由度では合計 $10^6 = 1000000$ 枚も必要となってしまう．また当然ながら，学習した物体しかうまく推定を行うことができない．

以上の問題を解決するため Yang らは，複数のカメラをわずかずつ位置姿勢を変化させて剛体結合したカメラシステムを用いる方法を示している [4]．そのシステムでは，ひとつのカメラを基準とし，複数のカメラを，それぞれが x, y, z 方向の微小な平行移動および各軸まわりの微小な回転に相当するカメラの運動を実現するように配置されている．それらのカメラの画像を元に，基準カメラが位置姿勢を変化させたときの基準カメラの画像変化を予測できるとすれば，カメラの位置姿勢の変化量が計算できるというのが，基本的な考え方である．カメラ間の位置姿勢の違いを微小なものとしておき，画像変化を線形近似することで，カメラの位置姿勢変化を連続量として算出する．彼らはこれを微分カメラ (differential camera) と呼んだ．

しかしながら，画像変化が，基準カメラの視点の周囲のわずか数個の視点から捉えた画像のみで表現できるというのは無理がある．Yang らは1次元画像と正射影かつカメラが平行移動するという仮定の下で，シーンの画像の濃淡分布を正弦波と見なして理論的な解析を行い，姿勢推定に一定の精度が保証される位置姿勢の範囲を算出するとともに，精度を担保するために画像をぼかすことを提案している．しかしながら現実には，カメラの回転運動や透視投影の考慮が必要であるし，シーンの画像の濃淡分布は連続的ではない．実際，われわれは Yang らの方法を実装し実験を行ってみたが，よい性能は得られなかった．少数のカメラの画像を元に，カメラの姿勢変化に伴う画像変化を線形近似することには，問題が多いとわれわれは考えている．

本研究では，以上の考察を元に，より多量のカメラを用いて現実的にカメラや物体の位置姿勢の変化を計算する方法を提案する．その方法は，Yang らと同じようにアピアランスベースであるが事前の学習を要しないものである．なお，問題を単純化して考えるため，カメラの前で物体を3自由度の回転を行わせたときに，その姿勢変化（あるいは回転速度）を画像から算出する問題を考える．具体的には，複数カメラをアレイ状に並べたカメラシステムを用いて，基準とする中心のカメラに対する見えを，周囲のカメラの以前の時刻における見えと直接マッチングすることで，姿勢変化を推定する．物体は，基本的には回転のみを行うものとするが，ある程度の位置の変化は許容される．本研究の動機の1つには，近年カメラの性能向上とともに値段が安価になってきており，多量のカメラをこのような目的に利用することが現実的になりつつあることがある．

2. カメラの動きと物体の動き

カメラが静止した状態で物体を R で回転させ $\boldsymbol{t}$ で平行移動させることと，物体が静止した状態でカメラを R^T で回転させ $-R^T\boldsymbol{t}$ で平行移動させることは，物体とカメラの相対運動として考えたとき同一の運動となる．したがって，物体の回転前のどのカメラ位置・姿勢から撮った画像が，物体の回転後の画像と同じになるのかを求めることができれば，そのときのカ

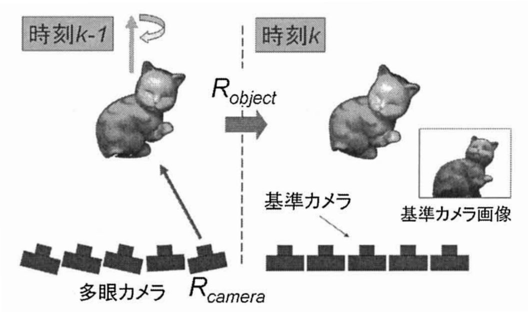

図 1　多眼カメラを用いた姿勢推定法の概要

メラ姿勢から物体の回転を計算することができる．

3. 多眼カメラを用いた姿勢推定法

推定を行う前処理として，対象となる物体のみが画像に写るようにするため背景差分を行う．以後取り扱う画像は全て背景差分後の画像とする．

3.1 推定法の概要

複数あるカメラの中から基準となるカメラを一つ選び，そのカメラを基準カメラと呼ぶこととする．図 1 をもとに本手法の基本的なアイデアを説明する．時刻 $k-1$ のときの複数あるカメラから得られた画像を $\boldsymbol{x}_i(i=1,2,\cdots)$ とし，時刻 k のときの基準カメラの画像を $\tilde{\boldsymbol{x}}_c$ とする．このとき，画像 $\tilde{\boldsymbol{x}}_c$ との類似度を最大とするような回転画像 $\hat{\boldsymbol{x}}_i(R)$ を求める．回転画像 $\hat{\boldsymbol{x}}_i(R)$ は，カメラ i が回転行列 R でその場回転をしたときの画像である．これはつまり，時刻 $k-1$ のときのどのカメラのどういう姿勢で撮った画像が画像 $\tilde{\boldsymbol{x}}_c$ に最も似ているかを求めていることになる．そして，類似度が最大となる回転画像を生成するカメラの姿勢 R_{camera} から，物体の画像フレーム間での姿勢変化 R_{object} を

$$\hat{R}_{object} = R_{camera}^T \tag{1}$$

と推定する．

ここで問題となるのは，

- 回転画像をどうやって作るのか
- 類似度をどう評価するのか
- 類似度が最大となる回転画像をどうやって求めるのか

の 3 点である．

3.1.1 回転画像の生成

2 次元射影変換によって，カメラをその場回転させたときの画像は作り出すことができる (図 2)．回転画像の座標 $\boldsymbol{m}_r = (u_r, v_r, 1)^T$ は元画像の座標を

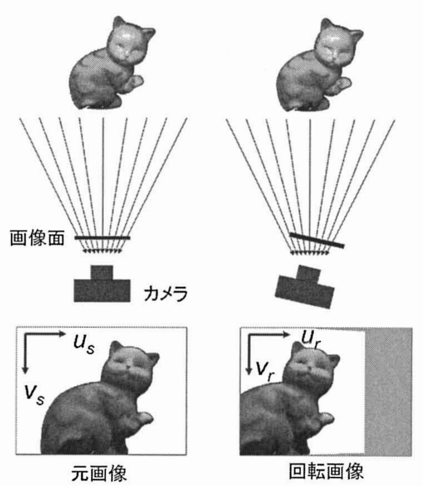

図 2　2 次元射影変換により，カメラをその場回転させた時の画像 (回転画像) を生成することができる．

$\boldsymbol{m}_s = (u_s, v_s, 1)^T$ とすれば

$$\boldsymbol{m}_r \propto AR^{-1}A^{-1}\boldsymbol{m}_s \tag{2}$$

となる．ここで A はカメラの内部パラメータである．

基準カメラで撮影された画像との類似度を計算するので，回転画像も基準カメラの内部パラメータ A_c で生成されたものにする必要がある．この新しい回転画像の座標を $\hat{\boldsymbol{m}}_r = (\hat{u}_r, \hat{v}_r, 1)^T$ とすると

$$\hat{\boldsymbol{m}}_r = A_c A^{-1}\boldsymbol{m}_r \tag{3}$$

$$\hat{\boldsymbol{m}}_r \propto A_c R^{-1}A^{-1}\boldsymbol{m}_s \tag{4}$$

となる．

3.1.2 類　似　度

画像 $\tilde{\boldsymbol{x}}_c$ と回転画像 $\hat{\boldsymbol{x}}_i(R)$ の類似度を評価する評価関数 J を

$$J = \frac{\|\tilde{\boldsymbol{x}}_c - \hat{\boldsymbol{x}}_i(R)\|^2}{n} \tag{5}$$

と定義する．ここで n は画像 $\tilde{\boldsymbol{x}}_c$ と回転画像 $\hat{\boldsymbol{x}}_i(R)$ の共通画素数である．共通画素とは図 3 において太枠で囲まれている領域のことである．J が小さいほど類似度が高いということになる．

3.1.3 類似度を最大とする回転画像の計算

評価関数 J を最小とするような画像 $\hat{\boldsymbol{x}}_i(R)$ を求める．まず，各カメラごとに J_{min} を計算する．これは非線形最適化問題となるが，ここでは最急降下法を用いて計算を行った．そして，全てのカメラの J_{min} を比較することで，画像 $\tilde{\boldsymbol{x}}_c$ に最も類似度が高い画像を生成するカメラ位置 i とそのときの回転行列 R を求

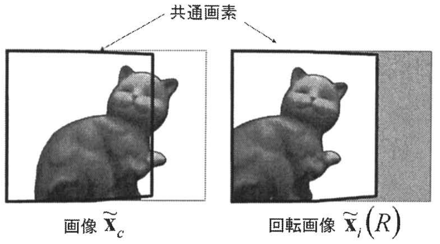

図 3　太枠で囲まれている領域が共通画素領域である．

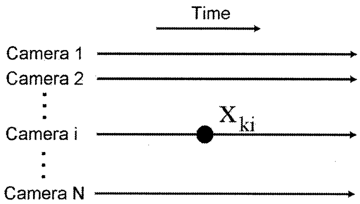

図 4　複数カメラの時系列画像を表す概念図．横軸は時間軸を表している．時刻 k のときのカメラ i の画像を $\boldsymbol{x}_{ki}$ とする

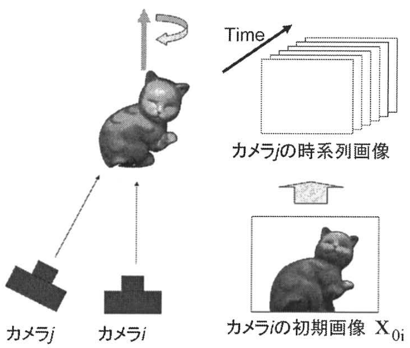

図 5　カメラ i の初期画像 $\boldsymbol{x}_{0i}$ と同じ画像がカメラ j の時系列画像の中に存在する．

めることができる．

3.2　推定精度の向上

前節までで述べた推定法をより一般化することにより，推定精度の向上を図る方法を述べる．

複数あるカメラの時系列画像を表す概念図を図 4 に示す．時刻 k のときのカメラ i の画像を $\boldsymbol{x}_{ki}$ とする．図 5 において，物体の回転速度に比べ，カメラのフレームレートが十分に高ければ，カメラ j の時系列画像の中にカメラ i の初期画像 $\boldsymbol{x}_{0i}$ と同一となるものが存在するはずである．このときのカメラの組み合わせ (i,j) およびその時間を見つけ出すことにより，推定精度の向上を図ることができる．5 × 5 のアレイ状にカメラが配置されていた場合，カメラの組み合わせ

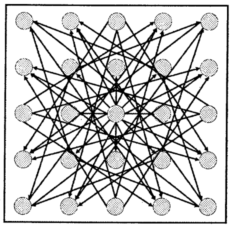

図 6　5 × 5 の例．選択できるカメラの組み合わせ

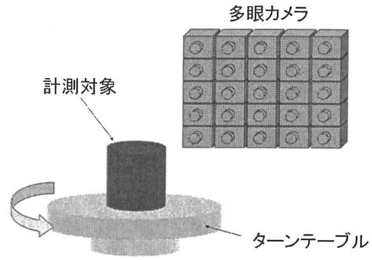

図 7　カメラの前にターンテーブルを配置し，計測対象を回転させたときの回転角度を推定する．

(i,j) の取り方は図 6 のようになる．選べるカメラの組み合わせの数が多いほど，回転方向の分解能をあげることができる．

類似度が最大となるカメラの組み合わせ (i,j) と，その時間を見つけ出す方法について述べる．まず，

$$J_{min(k,i,j)} = \min \frac{\|\boldsymbol{x}_{kj} - \hat{\boldsymbol{x}}_{0i}(j,R)\|^2}{n} \tag{6}$$

と定義する．ここで画像 $\hat{\boldsymbol{x}}_{0i}(j,R)$ は画像 $\boldsymbol{x}_{0i}$ を回転行列 R で回転させ，さらにカメラ j の内部パラメータに変換した回転画像のことである．$J_{min(k,i,j)}$ は，カメラ i の初期画像 $\boldsymbol{x}_{0i}$ とカメラ j の時刻 k のときの画像 $\boldsymbol{x}_{kj}$ との最適化された類似度をあらわしている．全てのカメラの組み合わせ，時間に対して $J_{min(k,i,j)}$ を計算することで，最も類似度が大きくなるカメラの組み合わせ (i,j) と時刻 k が求まる．

4.　実 験 結 果

本手法を有効性を検証する実験を行った．なお本実験では 3.2 節で述べた推定精度の向上は行っていない．

4.1　使用したカメラ

カメラはビュープラス社製の ProFUSION25 を使用した．5 × 5 で配置されたカメラのうち，中心にあるものを基準カメラとした．カメラの元の画像サイズは 640 × 480 画素だが，これを 80 × 60 画素に縮小して用いた．

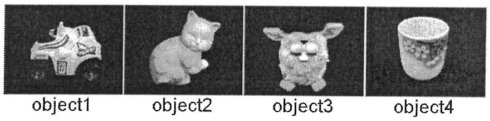

図 8　計測対象には，車のおもちゃ，ねこの置物，ぬいぐるみ，湯のみを使用した．

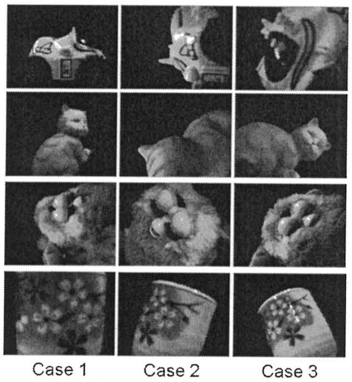

図 9　物体を回転させる前の基準カメラの画像．各計測対象において，3 つの異なるカメラ姿勢で実験を行った．

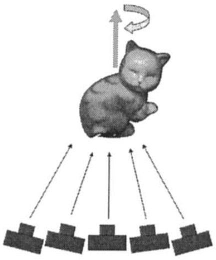

図 10　各カメラにおいて，類似度が最大となるカメラ姿勢は物体の回転角度によらずほぼ一定となる．

4.2 実　　験

ターンテーブルを使って計測対象を空間中のある軸回りで回転させ，そのときの姿勢変化を推定させる (図 7)．そして，推定された姿勢変化から軸回りの回転角度を算出し，実際の回転角度との比較を行った．この際，カメラの姿勢は適当に決めた．計測対象は図 8 に示したものを用いた．回転軸を変化させた場合を検証するため，各計測対象につき実験を 3 回行い，全て異なるカメラ姿勢とした．回転前の基準カメラの画像を図 9 に示す．

推定結果と推定誤差を図 11 に示す．この結果から，全ての計測対象，カメラ姿勢において，粗くではあるが回転角度が推定できていることが分かる．

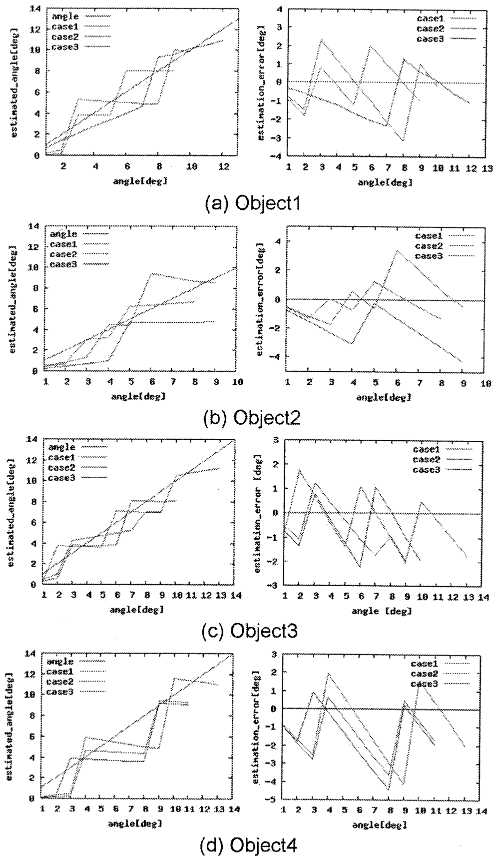

図 11　左側が推定角度，右側が推定誤差を示している．

推定角度が階段状となっている原因について考察する．物体が回転するとき，各カメラにおいて類似度が最大となる画像を生成するカメラ姿勢は物体の回転角度によらずほぼ一定となってしまう (図 10)．これにより，類似度が最大となる画像を生成するカメラが変化しないときは，推定角度もほぼ一定で，カメラが切り替わるときに推定角度も大きく変化してしまう．

カメラの前で腕を回転させたときのこぶしの姿勢変化を推定させた実験におけるスナップショットを図 12 に示す．上側が基準カメラの画像で下側が姿勢変化を示している．姿勢変化は，空間中の軸とその軸回りの回転で表し，回転角度を軸の長さで表した．腕が横向きのときと縦向きのときで，ほぼ正しい回転軸が推定できているように見える．

5. おわりに

アピアランスベースの推定法は，形状モデルや特徴

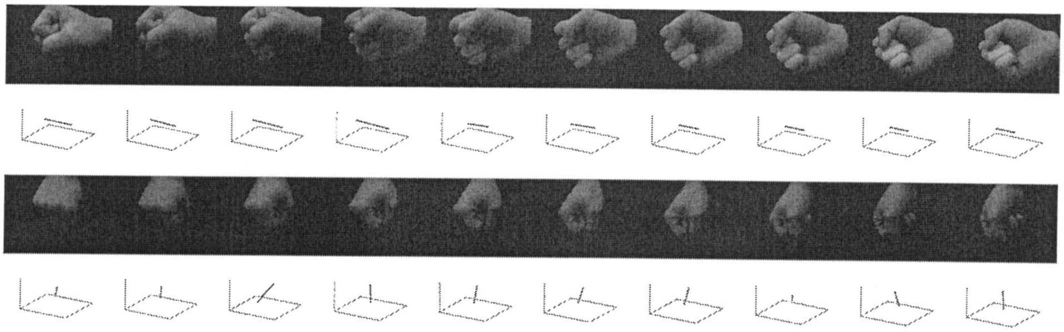

図 12　腕を回転させたときのスナップショット．上側が基準カメラの画像，下側が姿勢変化の推定結果．

点追跡が必要がないといった大きな利点が存在する．しかし，これまでに提案されている見えを学習する方法では，物体の自由度が増すにつれ学習用の画像を用意することが大変になることや，学習した物体以外はうまく推定することができなかった．

そこで本研究では，複数の固定カメラを用いることによって，事前の学習を必要としないアピアランスベースの姿勢推定を実現することを目指した．カメラが複数あることを活かして，物体の回転後の基準カメラ画像との類似度を最大とする回転画像を物体の回転前の複数のカメラから取得される画像から作り出し，それらを比較することで物体の姿勢変化を推定する手法を提案した．任意の物体の姿勢変化 (3 自由度) が推定できることを示した．

現在，類似度が最大となる画像はオフラインで計算を行っているが，GPU のプログラミングなどにより，リアルタイムでの推定を目指していきたい．

文　献

[1] P. Belhumeur and D. Kriegman: What is the set of images of an object under all possible illumination conditions, In *International Journal of Computer Vision*, 28(3), pp. 245-260, 1998

[2] K. Deguchi: A Direct Interpretation Dynamic Images with Camera and Object Motions for Vision Guided Robot Control, In *International Journal of Computer Vision*, 37(1), pp. 7-20, 2000

[3] H. Murase and S. Nayar: Learning and Recogition of 3D Objects from Appearance, In *Proc. of IEEE WorkShop on Qualitative Vision*, pp. 39-50, 1993

[4] H. Yang *et al.*: Differential Camera Tracking through Linearizing the Local Appearance Manifold, In *Proc. of IEEE Conference on Conputer Vision and Pattern Recogition*, pp. 1-8, 2007