カメラと物体の3次元運動を理解するための時系列シーンフローの分離

梅根健太[1]　酒井智弥[2]

概要：3 次元空間のシーンの変動を表現したシーンフローは，カメラの自己運動と物体の運動それぞれに起因する 3 次元ベクトル場の重ね合わせである．本研究では，与えられたシーンフローの時系列を自己運動と物体運動に起因するシーンフローに分離する手法を提案する．見かけのシーンフローは 2 次元画像上のオプティカルフローと深度画像のレンジフローとして推定される．深度の変化が無視できるならば，自己運動による見かけのシーンフローの時系列は，カメラの並進と回転のみの自由度に限定されるため，異なる時刻のシーンフローは少数の共通の基底ベクトル場の線形結合で表現可能である線形従属性をもつ．一方，運動する物体は時空間の一部に限定されるため，そのシーンフローはスパースである．ゆえに，線形従属性とスパース性をもつ時系列シーンフローの計算を，行列の低ランク＋スパース近似問題と見なし，これを凸最適化の方法で解く．この手法により，並進と回転の組合せによる自己運動の解釈や，運動物体の検知が可能となることを示す．

キーワード：画像認識・理解，　シーンフロー

Decomposition of Scene-Flow Sequence for Understanding 3D Ego and Object Motions

Kenta Umene[1]　Tomoya Sakai[2]

Abstract: The three-dimensional motion vector field in a scene is known as the scene flow, and it is composed of those fields due to camera egomotion and motion of objects. This paper presents the decomposition of a given temporal sequence of scene-flow fields into the sequences corresponding to ego and object motions. Because of the few degree of freedom by camera's translation and rotaion, the egomotional scene-flow fields can be assumed to be linearly dependent and expressed as linear combinations of a small number of basis vector fields. It can also be assumed that the scene flow caused by object motion is spatio-temporally sparse. The scene-flow decomposition is therefore posed as a low-rank and sparse matrix approximation problem, which is efficiently solvable by convex optimization techniques. Experimental results show that the proposed approach can help with combinatorial interpretation of motions.

1. はじめに

本研究は動画像中の見かけの動きから自己運動と物体運動を分離し，自己運動を解釈するための並進と回転運動の有無や動いている物体の検出を目的としている．見かけの動きは観測者 (カメラ) の動きとシーンを構成する物体の動きのふたつの原因によって生じる．これらの自己運動と物体運動の組合せにとして見かけの動きを解釈することで，カメラの並進と回転による典型的な自己運動や，動く物体の存在を検知できる．このような見かけの動きの解釈は，例えば光学センサを備えた車両の自動運転制御や障害物の検知のために必要である．

3 次元空間における物体の運動を表現した 3 次元ベクトルはシーンフローと呼ばれる．シーンフローは自己運動と物体運動によるものが混合した動きを表している．シーン

[1] 長崎大学工学部
Nagasaki University, 852–8521, Japan
[2] 長崎大学大学院工学研究科
Nagasaki University Graduate School of Engineering

© 2018 Information Processing Society of Japan

フローがもつ性質のみによってシーンフローを自己運動によるものと物体運動によるものへ分離できれば，厳密なカメラモデルや様々な物体に関する事前知識が不足していても，自己運動や物体運動を理解する手掛かりとなる．自己運動によるシーンフローは，カメラの並進と回転の高々 6 自由度の運動に起因している．一方，運動する物体の存在はシーンの一部の領域と時間帯に限られると仮定できる．自己運動の時系列シーンフローは限定された自由度の運動に起因するシーンフローの組合せであるため，複雑な自己運動によるシーンフローも並進と回転から生じるシーンフローの基底ベクトル場を用いた線形結合によって簡潔に表現できる．運動物体は時空間の一部にのみ存在するため，その運動に起因するシーンフローも時系列シーンフローの中でも一部しか存在しないというスパース性がある．見かけのシーンフローは，2 次元動画像と深度画像の時系列からそれぞれ推定されるオプティカルフローとレンジフローとして観測される．これらの性質を利用して，行列で表現した見かけの時系列シーンフローを，低ランクで表した自己運動に起因するシーンフローと，スパースである物体運動に起因するシーンフローへ分離するための最適化問題を解く．これにより，それぞれのシーンフローから自己運動の解釈や運動物体の検出ができる．

2. シーンフローの性質

シーンフローは 3 次元空間における物体の運動を 3 次元ベクトルで表現したものである [1]．見かけのシーンフローの第 1・第 2 成分は 2 次元動画像のオプティカルフロー [2] として観測される．第 3 成分は深度画像から推定されるレンジフローである．深度が一定ならば，自己運動に起因する見かけのシーンフローは，並進と回転の 6 自由度の運動によるシーンフローに限定される．以下，本稿では見かけのシーンフローを単にシーンフローと呼ぶ．

自己運動によるシーンフローとは無関係な物体の運動によるシーンフローがある．また，運動する物体は動画像の一部の領域と時間帯にのみ存在するため，そのシーンフローも時系列シーンフローの一部にしか表れない．物体運動によるシーンフローは，自己運動によるシーンフローからの外れ値であると言える．

自己運動に起因するシーンフローは深度の変化が無視できるならば，その x 成分，y 成分からなるオプティカルフローは図 1 のように並進と回転の運動によるオプティカルフローの組み合わせであると限定できるため，時系列シーンフローにおいて自己運動のシーンフローはカメラの並進と回転に起因するシーンフローを基底ベクトルとした線形結合として表現が可能である．

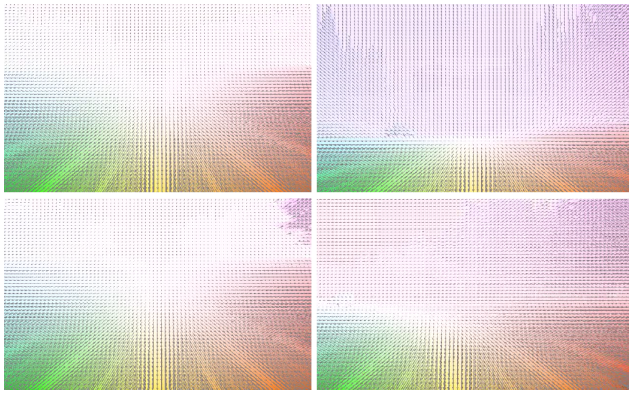


図 1 異なる時刻での見かけの動きによるオプティカルフロー．どの時刻においても自己運動のオプティカルフローは回転と並進の運動に起因するオプティカルフローで構成されている．

時系列シーンフローを行列で表現すると，自己運動に起因するシーンフローは基底ベクトル場を組み合わせた低ランクな行列として，外れ値である物体運動に起因するシーンフローはスパースな行列で表現ができる．時系列シーンフローを低ランク成分とスパース成分のシーンフロー行列に分離する手法としてロバスト主成分分析 (RPCA)[3] の誤差を考慮した SPCP(Stable Principal Component Pursuit) がある．

3. 低ランク成分とスパース成分への分離

自己運動と物体運動のシーンフローを分離するにあたり，時系列のシーンフローをまとめて 1 つの行列にする．図 2 に示すように，画素数が M=W×H の画像全体のシーンフロー $\boldsymbol{d} = (d_x,\ d_y,\ d_z) \in \mathbb{R}^3$ を x 成分，y 成分，z 成分ごとに $\boldsymbol{d}^{(0,0)}$ から順に $\boldsymbol{d}^{(W-1,H-1)}$ まで列ベクトルの形にすると，シーンフローの 3 成分をそれぞれ格納した M×1 の列ベクトルが 3 つできる．これらを縦に結合し，3M×1 のベクトルを構成する．このベクトルをデータが N 個ある時系列のシーンフローで構成する．N 個を結合することで 3M×N の時系列シーンフローを成分ごとに並べた一つの行列に変換する．この行列に対して分離の計算を行う．

© 2018 Information Processing Society of Japan

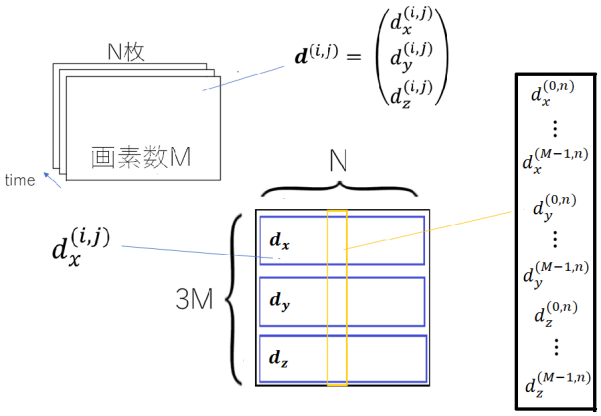


図 2 画素数 M の画像のシーンフローにおいて x 成分，y 成分，z 成分ごとにまとめた縦ベクトルをつくる．これを N 個あるデータすべてで行うことで時系列シーンフローを成分ごとにまとめた行列を構成する．

見かけの動きによるシーンフローの行列を $\mathbf{D}$，自己運動と物体運動のそれぞれに起因するシーンフロー行列を $\mathbf{L}$，$\mathbf{S}$ とする．ロバスト主成分分析では行列 $\mathbf{D}$ は式 (1) の最適化問題を式 (2) の条件のもとで解くことで行列近似される．

$$\min_{\mathbf{L},\mathbf{S}} ||\mathbf{L}||_* + \lambda ||\mathbf{S}||_G \tag{1}$$

$$subject\ to\ ||\mathbf{D} - (\mathbf{L} + \mathbf{S})||_F \leq \varepsilon \tag{2}$$

$||\cdot||_*$ は核ノルムを表しており，式 (3) で定義される．

$$||\mathbf{L}||_* = \sum_{i=1}^{min\{m,n\}} \sigma_i \tag{3}$$

$$||\mathbf{S}||_G = \sum_{j=1}^{N}\sum_{i=1}^{M} \sqrt{{s_x^{(i,j)}}^2 + {s_y^{(i,j)}}^2 + {s_z^{(i,j)}}^2} \tag{4}$$

σ_i は行列 $\mathbf{L}$ の特異値であり，式 (3) の最小化を行うことで自己運動に起因するシーンフローの行列の低ランク化を目指している．$\mathbf{S}$=(s_x，s_y，s_z) はスパース成分を表し，式 (4) で定義するように 3 次元ベクトルであるシーンフローの各成分に対応する要素を用いたグループ最小化によってシーンフローの 3 次元ベクトルのノルムが小さくなるような性質をもつ．

式 (2) の最適化問題は近接写像による最適化問題は交互方向乗数法 (ADMM) によって解くことができる [4]．

4. 実験

見かけの動きから自己運動と物体運動のそれぞれに起因するシーンフローを検出できることを確認する．
今回の実験では EISATS の SET2[5] のデータを使用する．このデータでは図 3 にあるように車載カメラから道なりに前進する自動車の前方を 640×480 の合成画像として 395 枚撮影しており, その時のシーンフローの真値は与えられている．

図 3 第 101 のフレームにおける車載カメラからの合成画像. 坂道を下りながら対向車とすれ違おうとしている．

実験に際して計算時間を短縮するために 640×480 のデータサイズを 160×120 に変更を行った．M=19200，N=395 であり，57600×395 の行列を構成した．分離計算における重み λ の値は 0.017，許容誤差 $\varepsilon = 0.01 \times ||\mathbf{D}||_F$ とした．分離によって得られた行列 $\mathbf{L}$ の左特異ベクトルはシーンフローの基底ベクトル場を表す．図 4 は最大の特異値に対応する基底ベクトル場である．

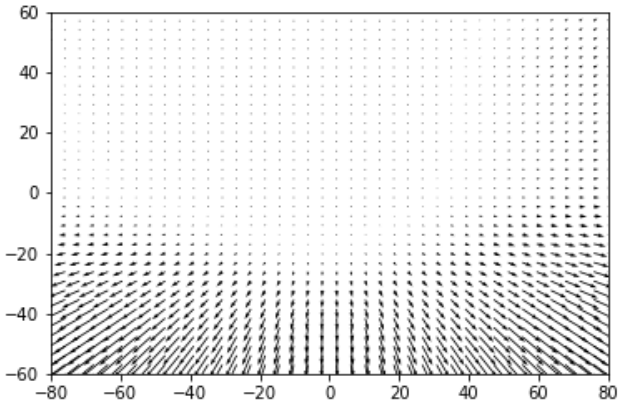


図 4 最も大きな特異値によって計算される，自己運動に起因するシーンフローを構成するベクトル場の一つ．

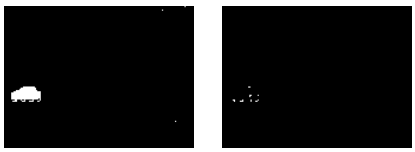
基底ベクトル場の x 成分と y 成分が示すオプティカルフロー場は放射状に広がっており，前進運動に起因するシー

© 2018 Information Processing Society of Japan

ンフローであると推測される．これはカメラが常に前進し続ける動画像のシーンフローに常に含まれる主成分である．画像の上半分のベクトル場と比べ下半分のベクトルが大きい理由は，画像の下部ほど路面の深度が浅いからである．

分離によって得られたスパース成分の行列 **S** から検出できる第 101 フレームにおける運動物体の位置を図 5 で示す．

(a) 正解画像 **(b)** $\lambda = \frac{1}{\sqrt{3M}}$ **(c)** $\lambda = \frac{2}{\sqrt{3M}}$



(d) $\lambda = \frac{4}{\sqrt{3M}}$ **(e)** $\lambda = \frac{8}{\sqrt{3M}}$

図 5 重み λ の値を変えて得られた第 101 フレームにおける運動物体の位置を表す二値画像．(a) は運動物体の正解の位置であり，(b) から (e) は λ を大きくしていったときの運動物体の検出結果を示す．

これらの検出の再現率，適合率，F 値をそれぞれ表 1 のようになった．

表 1 各 λ の値における物体検出の再現率と適合率および F 値

λ	$\frac{1}{\sqrt{3M}}$	$\frac{2}{\sqrt{3M}}$	$\frac{4}{\sqrt{3M}}$	$\frac{8}{\sqrt{3M}}$
再現率	0.98524	0.94834	0.90406	0.06642
適合率	0.19236	0.65228	0.98394	1.00000
F 値	0.32188	0.77293	0.94231	0.12457

λ の値が $\frac{4}{\sqrt{3M}}$ のときが最も性能の高い運動物体の検出が可能であった．適合率と比べ再現率が低い原因は，画像左中央に小さく映る運動物体の検出漏れである．深度の変化量が大きい対向車と比較して，カメラから遠い物体運動によるシーンフローが小さいことが検出漏れの原因であると考えられる．F 値より λ は $\frac{4}{\sqrt{3M}}$ 付近で運動物体の検出に最適な値をもつと予想される．

5. 結論

本研究は見かけの動きを表すシーンフローを自己運動と物体運動のそれぞれに起因するシーンフローに分離することで，自己運動の解釈や運動物体の検出を行うことを目的としている．自己運動での深度の変化が無視できるならば，カメラは並進と回転運動のみを行うため，その自由度は 6 である．このことから時間変化が起こった場合でも自己運動は共通の少数の基底ベクトル場を用いて表現することができる．また，今回使用したデータでは時空間で運動物体は一部しか存在しないので，物体運動に起因するシーンフローはシーンフローのデータ全体に対してスパースであると言える．これらを利用して SPCP を用いることで，自己運動と物体運動に起因するシーンフローをそれぞれ低ランク成分，スパース成分として分離する手法を提案した．

提案手法によって車載カメラから撮影した合成画像及び真値が与えらているシーンフローを用いて見かけの動きから自己運動と物体運動を分離し評価をした．行列 **S** がスパースとなるように重みを変えながら実験を行い，分離によって得られた自己運動に起因するシーンフローのうち，最大の特異値を使うことで，自己運動を最も構成している運動を表すシーンフローのベクトル場を得ることができた．スパース成分ではシーン中の運動物体の位置を正解画像と比較しながら再現率，適合率，F 値によって評価を行い，検出できたことを確認した．

実験では運動物体が時空間で一部のみ存在するデータを使用したため，その条件を満たさないデータでの検証は必要である．また，提案手法による分離を行った結果，それぞれの運動に起因するシーンフローが真値であるのかという確認も求められる．

今回は分離を行う際の重み λ の値は経験則的に決定したため，λ の最適化を行い値を調節することによってより正確に自己運動を構成するベクトル場と運動物体の位置を検出することが期待できる．

参考文献

[1] Vedula, Sundar and Baker, Simon and Rander, Peter and Collins, Robert and Kanade, Takeo: Three-dimensional scene flow, Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on, vol. 2, pp. 722–729, IEEE, 1999.

[2] Horn, Berthold KP and Schunck, Brian G: Determining optical flow, Artificial intelligence, vol.17, pp. 185–203, 1981.

[3] Candès, Emmanuel J and Li, Xiaodong and Ma, Yi and Wright, John: Robust principal component analysis?, Journal of the ACM (JACM), vol. 58, no. 3, pp. 11:1–11:37, 2011.

[4] Gabay, Daniel and Mercier, Bertrand: A dual algorithm for the solution of nonlinear variational problems via finite element approximation, Computers and Mathematics with Applications, vol. 2, no. 1, pp. 17–40, 1976.

[5] Vaudrey, Tobi and Rabe, Clemens and Klette, Reinhard and Milburn, James: Differences between stereo and motion behaviour on synthetic and real-world stereo sequences, in 23rd International Conference of Image and Vision Computing New Zealand (IVCNZ '08), 2008, pp. 1–6.

© 2018 Information Processing Society of Japan