複合現実感研究会 MR2022-14
(2022.10. 6・ 7)

映像作品のカメラワーク数値化に向けた カメラと被写体の位置姿勢推定法

坂井 甚太[*1]　　宍戸 英彦[*2]　　北原 格[*2]

A Method for Estimating Position and Posture between Camera and Subject for Quantifying Camera-Work of Shooting Video

Jinta Sakai[1] , Hidehiko Shishido[*2] and Itaru Kitahara[*2]

Abstract --- Since camera work is positioned as an important technique that emphasizes the atmosphere and impression of a scene, it has attracted attention as an important element of video analysis. In this paper, we propose a method to estimate the position and posture of the subject and camera for camera work quantification. By applying Structure from Motion (SfM) to monocular video, the 3D information of the scene and the camera position and posture were obtained. The 3D human pose estimation was also used to estimate the posture of the subject from the images. We applied the proposed method to videos of dynamic scenes and evaluated the accuracy of estimating the position and posture of the subject and camera.

Keywords: camera work, 3D pose estimation, human pose, camera pose

1 はじめに

ディジタル放送，VOD（ビデオ・オン・デマンド），動画共有サービスの普及により，数多くの映像作品が日々制作されている．映像作品数の増加によって，映像への効率的なアクセス方式の需要が急激に高まりつつある．映像情報の解析に基づく映像データベース構築は，アクセス方式の効率化の重要なキーテクノロジとして注目を集めている．映像には，画像や音声といったパターン情報，撮影シーン中の物体の形状や色などの物理情報，それらによって表現される出来事や情景といった意味情報といった多様な情報が含まれている．パターン情報からの物理情報検出，物理情報に基づいた意味情報推定と分析処理が高度化するに伴い，データベースの分類性能が向上する．例えば，物理情報の検出に基づくデータベース構築では，映像の基本単位であるショット切り替え[1]や映像のカメラワーク[2][3][4]などの映像特徴によってデータベース分類が行われる．意味情報に基づく構築では，映像のインデクス[5]，要約[6]，キャプション[7]などより高次な映像特徴によってデータベースが分類できる．

我々は，意味情報の推定の手がかりである物理情報のうち，カメラワークに注目している．映画やテレビで被写体を撮影する際のカメラの操作法であるカメラワークは，効果的な場面表現をするための重要な演出技法である．カメラワークには，その撮影法ごとに決められた印象効果が存在するため，映像からのカメラワーク推定によって映像製作者が映像に込めた“演出”という，高次な意味情報を推定することができ，その結果高品質なデータベースの構築が期待できる．しかし，従来のカメラワーク推定は，カメラの動きのみに焦点を当て，被写体については考慮していない．その結果，一部の物理情報を欠落した状態で意味情報を推定することとなる．我々は，撮影シーンに含まれる多様な物理情報に基づいてカメラワークを推定することにより，意味情報の推定精度の向上を目指している．本稿では，撮影シーンの物理情報としてカメラと被写体の位置姿勢を取り上げ，被写体を含んだ映像シーケンスからカメラワーク推定法を提案する．

図 1 に提案手法の処理の流れを示す．映像作品（カメラワークが施された単眼映像）をフレーム分割し画像群を取得する．各フレームにおいて被写体の3次元骨格を推定する．被写体骨格の各関節の3次元座標と画像上での観測座標の対応関係に基づき PnP 問題[8]を解くことにより，カメラと被写体の位置関係を推定する．フレーム分割した画像群に Structure from Motion（SfM）[9]に代表される3次元フォトグラメトリを適用し，撮影シーンの3次元形状復元すると同時にカメラの位置姿勢を推定する．推定したシーンの3次元形状と骨格モデルを統合することにより，撮影シーンにおけるカメラと被写体の位置姿勢を推定する．

*1 筑波大学　システム情報工学研究群
*2 筑波大学計算科学研究センター
*1 University of Tsukuba
*2 Center for Computational Sciences, University of Tsukuba.

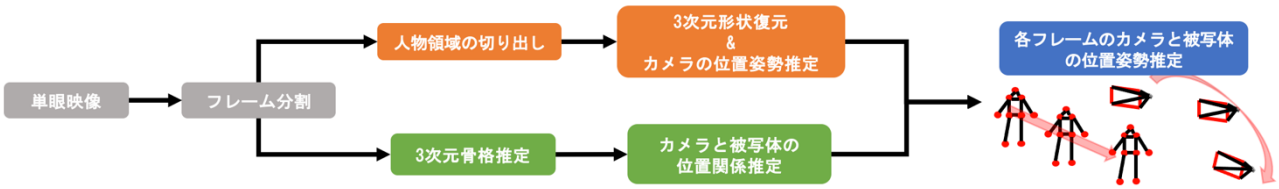

図1　単眼映像より分割した画像群から推定した3次元骨格を用いて計算したカメラと被写体の位置関係と，人物領域を切り出した画像から3次元形状復元することで推定した撮影シーンにおけるカメラの位置姿勢を用いて，撮影シーンにおけるカメラと被写体の位置姿勢を推定する．

Fig.1 The positional relationship between the camera and the subject calculated using the 3D pose estimated from the monocular video, and the position and posture of the camera in the shooting scene estimated by 3D shape restoration using images of cropped the human region are used to estimate the position and posture of the camera and subject in the shooting scene.

2　関連研究

カメラワークの推定手法の一つとして，Ewerth ら[2]は，MPEG ビデオから動きベクトルを抽出しフレーム間で特徴点の対応づけを行うことでカメラワークを解析する手法を提案した．この手法では，カメラの動作を静止，パン，チルト，並進と区別することが可能である．しかし，ズーム操作とカメラを前後に移動させる操作を区別して検出していない．また，この手法はカメラ操作に起因する以外の映像変化はないと仮定している．そのため被写体が運動した場合，カメラワークを正しく推定することが困難であるという問題がある．

Chong ら[3]は，複数のフレームから時空間投影画像を生成し，時間変化を解析することでカメラワークを推定する手法を提案した．この手法では，フレーム内の任意の直線部分の時間的変化が解析可能であり，輝度の変化が激しい映像に対しても映像内の動きを反映した特徴抽出が容易である．Yoshitaka ら[4]は，時空間投影画像生成時に動物体による影響が少なくなるよう直線を動的に決定することで，被写体が運動している映像からカメラワークの推定を行なった．しかし，Chong ら[3]や Yoshitaka ら[4]の手法では被写体を排除してカメラの動きを推定しており，被写体とカメラ相互の動きとしてカメラワークを推定していない．その結果，動く被写体に対してカメラを移動させる撮影技法と静的な被写体に対してカメラを移動させる撮影技法の区別が困難となっている．本研究では，従来手法によって推定されるカメラワークからは読み取りが困難な演出の検出や撮影術の抽出を実現するために，カメラの運動以外の要素である被写体の3次元位置姿勢も推定し，それらに基づいたカメラワーク推定を目指す．

3　映像作品のカメラワーク

3.1　映像作品から獲得できる情報

コンピュータビジョン技術を用いて映像作品(単眼映像)から取得可能な物理情報としては，Structure from Motion (SfM) [9]による“シーンの3次元点群”と“カメラの位置姿勢情報”，Semantic Segmentation [10]による“領域の属性情報”，骨格推定 [11]による“被写体の姿勢情報”が考えられる．

提案手法において撮影シーン中のカメラと被写体の位置姿勢を推定するためには，カメラの位置姿勢情報，およびカメラに対する人物の位置姿勢情報が必要である．撮影シーンの3次元点群とカメラの位置姿勢情報は SfM により取得可能であり，カメラに対する人物の姿勢は骨格推定により取得可能である．しかし，深層学習による単眼画像からの骨格推定では，人物サイズとカメラパラメータが既知でない限り人物の奥行き情報の取得が困難である．そのため，提案手法では人物サイズとカメラの内部パラメータは事前に取得可能な状況を対象とし，人物サイズとカメラの内部パラメータを用いてカメラに対する人物の位置を推定する．

3.2　カメラワークの種類

カメラワークは大きく以下の5種類に分類される．

(A)　固定撮影：カメラを固定したまま撮影する技法
(B)　パン・チルト：カメラを水平方向または垂直方向に回転させる撮影技法
(C)　ズーム：焦点距離の変化により被写体の見た目を拡大/縮小する撮影技法
(D)　移動撮影：シーン中の被写体に合わせてカメラを移動させながら撮影する技法
(E)　フォーカス：焦点距離の変化によりピントの合う場所を変更する撮影技法

(A)(C)(E)はカメラの位置姿勢は固定されており，(B)はカメラの姿勢は変化するが位置は固定である．その結果，(A)(B)(C)(E)で撮影された映像のフレーム間では運動視差が生じない．SfM は多視点画像間での特徴点の対応関係からその3次元情報を復元するため，運動視差の生じない場合，3次元復元が困難である．本稿では，運動視差の発生する(D)で撮影された映像を対象とした，カメラと被写体の位置姿勢情報取得法について述べる．

4 カメラと被写体の位置姿勢推定

本節では，単眼映像から各フレームのカメラと被写体の位置姿勢を推定する手法について説明する．

4.1 被写体を含む単眼映像からの3次元形状復元

カメラを移動しながら撮影した映像をフレーム毎に分割する．分割した画像群に Semantic Segmentation を適用し人物領域を検出する．SfM は画像群から検出した特徴点に基づいて3次元情報を推定するが，特徴点は静的な物体上に存在することを想定しているため，動的な被写体領域から検出された特徴点は，3次元復元誤差の原因となる．そこで人物領域を切り出し，黒く塗り潰す（輝度変化をなくする）ことで特徴点が検出されにくくする．

4.2 被写体とカメラの相対的位置関係の推定

フレーム毎に深層学習を用いて3次元骨格を推定する．透視投影画像が有するスケール不変性により，2次元画像から3次元骨格の実スケールを推定することは困難である．ここでは，被写体の特定部位のサイズなど事前収集情報を用いて，推定3次元骨格が世界座標系における被写体と同スケールになるようスケーリング処理を施す．スケーリングした骨格とその画像上での観測座標のペアおよび内部カメラパラメータに基づいて PnP 問題 [8]を解くことにより，人物座標系におけるカメラの位置を推定する．

4.3 3次元形状復元モデルと人物骨格の統合

4.2 節で取得した人物座標系におけるカメラの位置関係と SfM で求めた世界座標系におけるカメラの位置姿勢の対応関係から，カメラ座標系を介して人物座標系と世界座標系の間の剛体変換を求め，被写体の骨格座標を世界座標系に変換する．

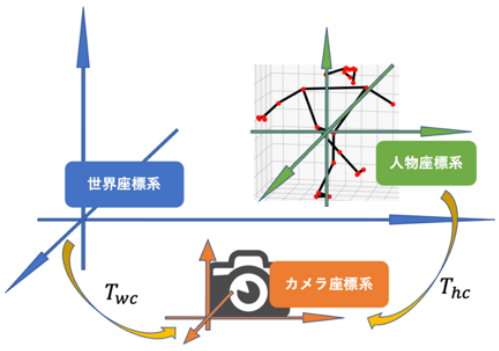

図2　世界座標系，カメラ座標系，人物座標系の関係．(世界座標系:SfM で取得，人物座標系:骨格推定で取得)

Fig.2 Relationship between world coordinate system, camera coordinate system, and human coordinate system. (World coordinate system: obtained by SfM, human coordinate system: obtained by 3D pose estimation)

図 2 に各座標系の関係を示す．世界座標系とカメラ座標系の剛体変換をT_{wc}，人物座標系とカメラ座標系の剛体変換をT_{hc}とする．人物座標系での人物の座標をX_h，世界座標系での人物の座標をX_wとすると，式(1)が成り立つ．

$$T_{wc} * X_w = T_{hc} * X_h \quad (1)$$

式(1)を式(2)に変形することで，世界座標系における被写体（人物）の位置姿勢が求められる．

$$X_w = {T_{wc}}^{-1} * T_{hc} * X_h \quad (2)$$

5 評価実験

本節では，カメラと被写体の位置姿勢推定精度の測定に関する評価実験について述べる．

5.1 提案手法の実装

4.1 節で説明した Semantic Segmentation として Mask R-CNN に基づく手法 [12]を適用し人物領域を検出する．4.2 節で説明した骨格推定として，深層学習に基づいて人体のキーポイントを推測する BlazePose [13]を採用し，フレーム毎の被写体(人物)の3次元骨格を推定する．撮影シーンの3次元形状復元には，SfM に基づく写真測量ソフトウェアである Pix4Dmapper [14]を用いた．撮影シーンは，3D 制作プラットフォームである Unreal Engine 5 を用いて 3DCG を制作するため，カメラと被写体の正確な位置姿勢を取得可能である．撮影（CG レンダリング）時のカメラパラメータは，35mm フルサイズイメージセンサ，焦点距離 50mm のカメラと同等の設定とした．被写体となる人物の身長は 160cm とした．

5.2 カメラワークによる人物およびカメラの推定精度

カメラワークによる人物およびカメラの推定精度の確認を目的とした評価実験を実施した．フレームをコマ撮りする毎に，フレーム間のカメラの移動距離，および各フレームでのカメラと被写体の距離を取得し，位置推定の正解データとする．姿勢の正解データは，フレーム間のカメラの姿勢変化量や各フレームのカメラ姿勢における人物の姿勢を計算し取得する．

本実験で対象とした映像を撮影するカメラの移動は，以下の4種類に分類される．

(1) ドリーイン：カメラが被写体に近づく
(2) ドリーアウト：カメラが被写体から遠のく
(3) トラック：動く被写体に対してカメラ自体を追随させる
(4) アーク：カメラが被写体を中心に回り込む

(1)(2)(3)(4)の撮影法で撮影した各フレームの画像を図 3 に示し，撮影した状況を図 4 に示す．(1)の撮影では，

被写体から 11m 離れた場所からカメラを 1m ずつ近づけ，2m の距離まで近づけた．(2)の撮影では，被写体から 2m 離れた場所からカメラを 1m ずつ遠ざけた．(3)では，前進する被写体に対しカメラを後退させて撮影した．カメラと被写体の距離を4m に固定し，1m ずつ移動させた．(4)では，中心に位置する被写体に対し，カメラを 10 度ずつ回り込むように移動させた．

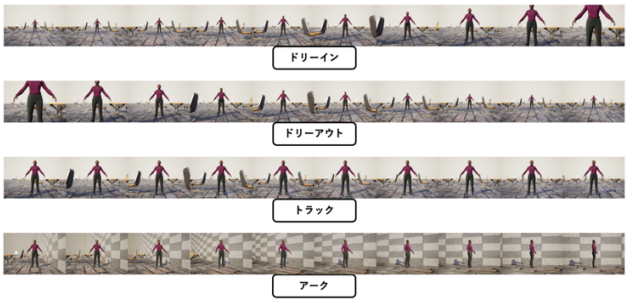

図3　各移動撮影で撮影した映像のようにコマ撮りした画像

Fig.3 Time-lapse images like the video captured

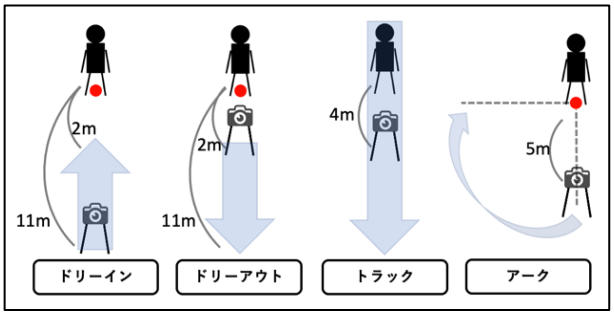

図4　移動撮影による撮影状況 ((1)ドリーイン (2)ドリーアウト (3)トラック (4)アーク)

Fig.4 Shooting situation ((1)Dolly in (2)Dolly out (3)Track (4)Arc)

図 5 にフレーム間のカメラの移動距離の誤差を示す．連続する2フレームに対し，カメラの座標間距離を算出し，真値と比較した．全ての撮影法において誤差の平均が 14cm 以内となった．(1)(2)(3)の誤差が(4)と比較し大きい原因は，被写体とカメラの距離が近いフレームで，背景から特徴点を検出することが困難であったことであると考えられる．

図 6 に各フレームのカメラと人物の距離の誤差を示す．カメラ座標系と人物座標系の原点間距離を算出し，真値と比較した．全ての撮影法において真値よりも距離が大きく推定され，(1)(2)では平均が 50cm を超える誤差となった．カメラと人物の距離は推定した骨格座標を基に PnP 問題を解くことで推定しているため，推定した骨格座標に誤差が生じた場合，誤差の伝播が起きる．今回使用した BlazePose は人物を遠くから撮影した LSP データセットを用いて学習されている．そのため骨格座標をカメラ座標に投影する際，視点の位置が無限遠に存在する平行投影が用いられていると考えられる．平行投影は透視投影と比較すると遠近感が少なく，画像中に映る被写体の大きさが同じ場合，実際のカメラと被写体の距離は平行投影の方が遠くなる．本実験では，カメラと被写体の距離は 2-11m と近い距離で撮影されているため，平行投影を行うことで誤差が生じ，その誤差が伝播した結果誤差が生じたと考えられる．

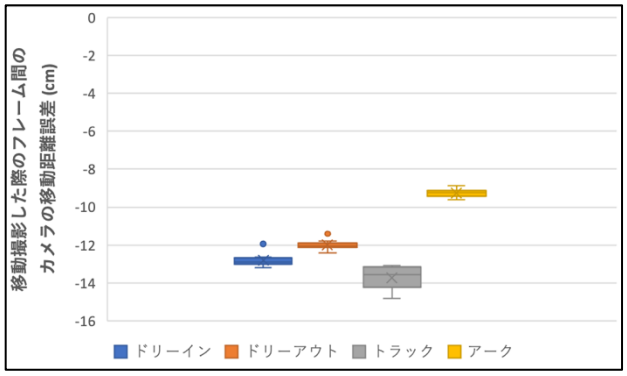

図5　フレーム間のカメラの移動距離誤差の平均

Fig.5 Average camera move distance error between frames

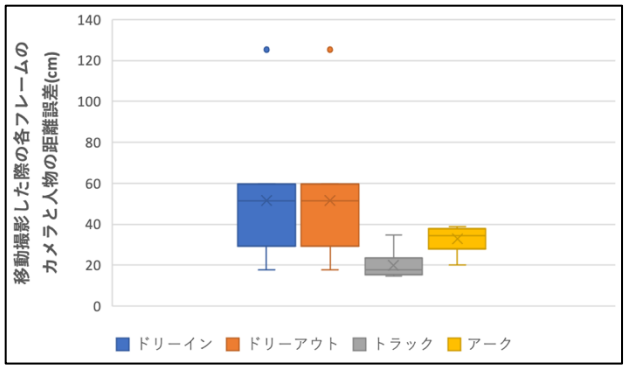

図6　カメラと人物の距離誤差の平均

Fig.6 Average of the distance error between the camera and the subject in each frame

図 7 にフレーム間のカメラの姿勢変化量の誤差を示す．連続する2フレームに対し，カメラ座標系の XYZ 軸方向それぞれの角度変化量を算出し，真値と比較した．全ての結果において，1 度以内の誤差となった．

図 8 と図 9 に各フレームのカメラ姿勢における人物姿勢の誤差を示す．図 8 の結果では Y 軸の誤差ならびに Z 軸の誤差が X 軸の誤差に比べて大きくなっている．つまり，人物が実際よりも人物座標系の X 軸方向に回転しており，前傾や後傾しているといえる．図 9 では Y 軸の誤差に比べて X 軸の誤差と Z 軸の誤差が大きくなっている．つまり，人物が実際よりも人物座標系の Y 軸方向に回転しているといえる．この誤差は骨格推定の際の平行投影が要因であると考えられる．人物の奥行き方向の誤差は人物座標系における Z 軸方向の回転には起

因しない．そのため，推定した人物の姿勢は，人物座標系の Z 軸に比べて X 軸と Y 軸方向に回転していると考えられる．

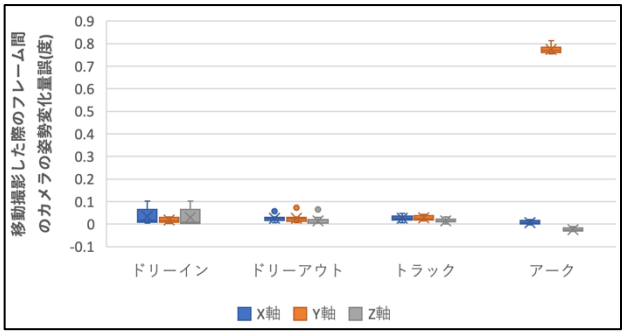

図7　フレーム間のカメラの姿勢変化量の誤差の平均

Fig.7 Average of the errors in the camera's posture change between frames

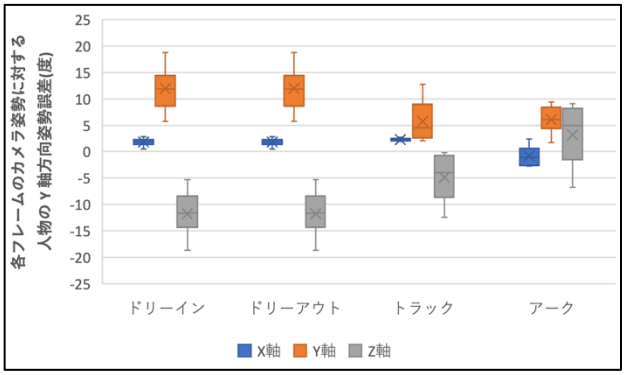

図8　カメラ姿勢に対する人物の Y 軸方向姿勢誤差の平均

Fig.8 Average of the errors in the camera's posture change between frames

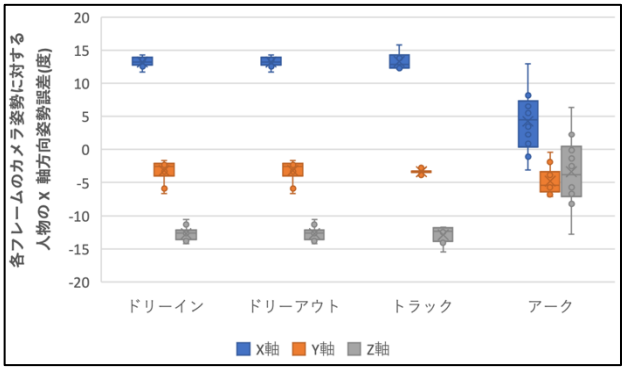

図9　カメラ姿勢に対する人物の X 軸方向姿勢誤差の平均

Fig.9 Average of the errors in the camera's posture change between frames

以上の結果より，撮影に影響を及ばさないと考えられる誤差内で，カメラの位置姿勢が推定できた．一方で，人物の位置推定は，既存の骨格推定モデルを用いると 50cm 以上の誤差，姿勢推定においては 10 度程度の誤差が生じた．

5.3 画像位置による人物の位置推定精度

ここでは，被写体が画像上で観測される位置がカメラと被写体の距離推定精度に与える影響について調査する．

5.3.1 画像位置による位置推定誤差

被写体の画像位置が移動するよう撮影した画像を図 10 に示す．被写体と人物の距離は 4m とした．図 10 の画像に対し被写体の骨格を推定し，推定した骨格情報を用いて PnP 問題を解くことでカメラと被写体の距離を求める．推定した距離と撮影した実際の距離を比較することで精度評価を行う．

図10 カメラを回転させることで被写体の写る位置をずらした画像

Fig.10 Images in which the subject's position is shifted by rotating the camera

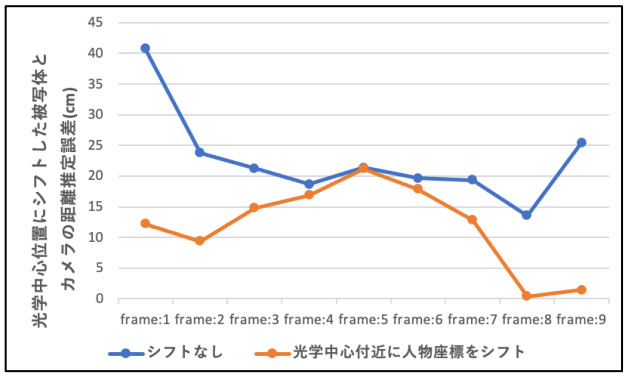

図11 被写体位置におけるカメラと被写体の距離推定誤差

Fig.11 Error in estimating camera-subject distance due to subject position

図 11 に被写体位置におけるカメラと被写体の距離推定誤差を示す．画像中心付近では，カメラと被写体の推定距離に大差がないが，画像端では距離推定精度が大きく下がった．5.2 節でカメラと被写体の距離推定誤差は投影法が原因と考えられると述べたが，画像端の方が画像中心よりもその影響が受けやすいと考えられる．

5.3.2 光学中心補正を追加した位置推定

5.3.1 節より被写体の写る画像位置が光学中心から離れている場合，カメラと被写体の位置関係の推定精度に影響を与えることが確認できた．そこで本節では，被写体の画像座標を光学中心付近にシフトして位置推定をした結果について述べる．

骨格推定時に人物座標系の原点とされている被写体

の腰の画像座標をH_{uv_h}，光学中心の座標をC_{uv}，シフト前の人物の各骨格座標をH_{uv}，シフト後の人物の各骨格座標をH'_{uv_h}とすると，シフト後の人物の各骨格座標は式(3)で求まる．

$$H'_{uv\_h} = H_{uv} + (C_{uv} - H_{uv\_h}) \quad (3)$$

図 11 に光学中心位置にシフトした被写体とカメラの距離推定誤差を示す．光学中心位置から離れている被写体ほど画像座標をシフトすると距離推定誤差が小さくなった．光学中心付近に被写体をシフトしたことにより，PnP 問題を解く際，カメラ座標系から投影面までの距離が小さくなり，推定距離も小さくなったと考えられる．元画像位置が光学中心から遠いほど，その効果は大きく現れる．また他の要因としては，光学中心付近に被写体をシフトしたことによる投影誤差の軽減が考えられる．

6 おわりに

本研究では，映像作品の演出情報を検出することを目的としたカメラワークの推定法において，カメラワークはカメラと人物の位置姿勢から推定できると考え，撮影シーンの3次元形状復元と人物の骨格推定を組み合わせることにより，撮影シーンにおけるカメラと被写体の位置姿勢を推定する方法を提案した．評価実験の結果，SfM を用いたカメラの位置推定精度は，フレーム間で 15cm 程度の誤差に収まることが確認できた．また，カメラの姿勢に関しては，フレーム間で 1 度以内の誤差になることを確認した．一方，人物の位置推定精度は，少なくとも 20cm 程度の誤差が生じ，最大では 50m を超える誤差となった．人物の姿勢推定においても 20 度程度の誤差が生じており，既存の骨格推定モデルは，直接適応することが困難であることがわかる．また，画像位置による人物の位置推定誤差は，光学中心に画像座標をシフトすることにより軽減した．

今後の課題としては，推定したカメラと被写体の位置姿勢を用いたカメラワークの推定法の考案や，新たな骨格推定モデルの生成による人物の位置姿勢推定精度の向上があげられる．

参考文献

[1] H. Liu, T.-H. Tan, and T.-Y. Kuo, “A Novel Shot Detection Approach Based on ORB Fused With Structural Similarity,” IEEE Access, vol. 8, pp. 2472–2481, 2020, doi: 10.1109/ACCESS.2019.2962328.

[2] R. Ewerth, M. Schwalb, P. Tessmann, and B. Freisleben, “Estimation of arbitrary camera motion in MPEG videos,” in Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004., 2004, vol. 1, pp. 512-515 Vol.1, doi: 10.1109/ICPR.2004.1334181.

[3] C.-W. Ngo, T.-C. Pong, H.-J. Zhang, and R. T. Chin, “Motion characterization by temporal slices analysis,” in Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No.PR00662), 2000, vol. 2, pp. 768–773 vol.2, doi: 10.1109/CVPR.2000.854952.

[4] 吉高淳夫, 松井亮治, and 平嶋 宗, “カメラワークを利用した感性情報の抽出,” 情報処理学会論文誌, vol. 47, no. 6, pp. 1696‐1707, 2006.

[5] Y. Li, S. Narayanan, and C. C. J. Kuo, “Content-based movie analysis and indexing based on audiovisual cues,” IEEE Trans. Circuits Syst. Video Technol., vol. 14, no. 8, pp. 1073–1085, 2004, doi: 10.1109/TCSVT.2004.831968.

[6] A. Yoshitaka and Y. Deguchi, “Video Summarization based on Film Grammar,” in 2005 IEEE 7th Workshop on Multimedia Signal Processing, 2005, pp. 1–4, doi: 10.1109/MMSP.2005.248620.

[7] L. Zhou, Y. Zhou, J. J. Corso, R. Socher, and C. Xiong, “End-to-End Dense Video Captioning with Masked Transformer,” in 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 8739–8748, doi: 10.1109/CVPR.2018.00911.

[8] X.-S. Gao, X.-R. Hou, J. Tang, and H.-F. Cheng, “Complete solution classification for the perspective-three-point problem,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 25, no. 8, pp. 930–943, 2003, doi: 10.1109/TPAMI.2003.1217599.

[9] J. L. Schönberger and J.-M. Frahm, “Structure-from-Motion Revisited,” in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4104–4113, doi: 10.1109/CVPR.2016.445.

[10] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3431–3440, doi: 10.1109/CVPR.2015.7298965.

[11] M. S. V. L. Bugra Tekin Isinsu Katircioglu and P. Fua, “Structured Prediction of 3D Human Pose with Deep Neural Networks,” in Proceedings of the British Machine Vision Conference (BMVC), Sep. 2016, pp. 130.1-130.11, doi: 10.5244/C.30.130.

[12] K. He, G. Gkioxari, P. Dollár, and R. Girshick, “Mask R-CNN,” in 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2980–2988, doi: 10.1109/ICCV.2017.322.

[13] V. Bazarevsky, I. Grishchenko, K. Raveendran, T. L. Zhu, F. Zhang, and M. Grundmann, “BlazePose: On-device Real-time Body Pose tracking,” ArXiv, vol. abs/2006.10204, 2020.

[14] “Pix4Dmapper,” http://www.pix4d.com/jp/product/pix4dmapper-photogrammetry-software