I-013

自由視点画像と顔姿勢推定を用いた被写者視点画像の生成

Object Person's View Generation Using Head Pose Estimation from Free Viewpoint Images

山崎 颯大† 北條 海斗 † 青木 輝勝†
Sota Yamazaki Kaito Houjho Terumasa Aoki

1. はじめに

本稿では，写真に写っている被写体が「何を見ているのか」を画像として出力する被写体視点画像システムを提案する．対象者の行動を見ることのできるシステムに監視カメラシステムがある．しかしながら，不審者を外から見ても本当に怪しい人間であるのかを特定することは困難な場合が少なくない．逆にもしも不審者が「今何を見ているのか」という情報を得ることができれば，犯罪抑止につながると考えられる．そこで本稿では自由視点画像を入力とし，被写体の視線の先を画像として出力する被写体視点画像システムを提案する．

自由視点画像生成技術の歴史は長いが，近年深層学習の導入により飛躍的に精度が向上した．また，従来多数（場面によっては 100 台単位）のカメラの設置が必要であったが，その数も大幅に低下し，誰もが自由視点画像を容易に生成できる環境が整いつつある．このような状況のもと，顔姿勢推定の技術を用いた被写者視点画像の生成とその応用について提案する．

本技術が利用可能になれば，任意の人物の視線を容易に見ることが可能になる．そのため，不審人物の特定に加え，講義中における生徒の集中度の確認，会議中でのいねむりや必要のないタブレット端末操作の特定，スポーツにおいて選手の視線から入手する情報の特定など幅広い分野で利用が可能となる．

本稿の構成は次の通りである．2.では監視カメラシステムの既存研究と近年の自由視点画像生成について概説する．次の 3.では提案手法について示す．続く 4.では，実験結果について説明し，最後に 5.では結論を述べる．

2. 既存研究

対象者の行動を見ることのできるシステムに監視カメラシステムがある．現在，監視カメラシステムに深層学習の手法を用いたものとして PV-YOLO[1]が提案されている．PV-YOLO は監視カメラの入力画像に対して物体検知(ObjectDitection)を取り入れた手法である．しかし，これらの手法では特に屋内環境を考慮した場合，カメラと対象者の間に遮蔽物(オクルージョン)が存在してはならないといった制約がある．したがって，意図的に第三者が遮蔽物を設置した場合，既存の監視カメラシステムでは対応できない．

近年，深層学習を導入することにより自由視点画像の生成の精度が飛躍的に向上した．深層学習を利用した自由視点画像を生成する手法の一つにニューラル場を用いた NeRF(Neural Radiance Fields)[2]がある．NeRF はネットワーク内部で光線空間を生成する手法である．入力に光線(Ray)の位置情報(x, y, z)と方向(θ, φ)から MLP ネットワークを用いて色情報と密度情報を予測し，光線空間の生成を行うことで高精度な自由視点画像の生成を可能にした手法である．

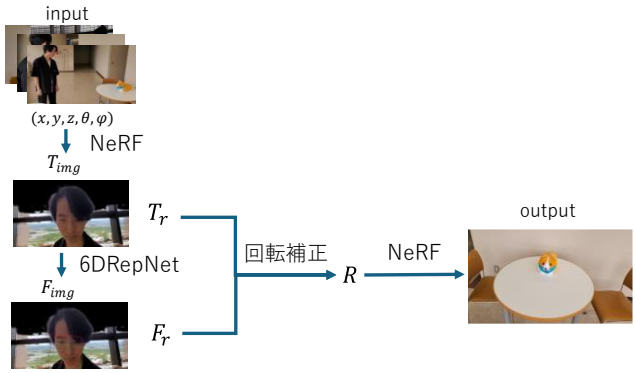

図 1 提案手法の流れ

本研究では，監視カメラシステムに自由視点画像の技術を取り入れ，上述のオクルージョン問題を解決するとともに対象者の視線を容易に確認できる被写体視点画像生成システムを構築する．

3. 提案手法

自由視点画像から顔の正面の画像T_{img}を抽出する．T_{img}に対応するカメラ行列T_rと顔姿勢推定から得られた被写体の顔の向き対応するカメラ行列F_rの回転ベクトルの向きを一致させる処理を行うという流れである．本稿の提案手法像に対して顔姿勢推定をかけその座標を光線空間に反映させることによって被写体視点の生成を行う．被写体視線の生成までの流れを図 1 に示す．

まず，自由視点画像生成は NeRF を用いて，数十のカメラからニューラル乗の生成を行う．NeRF では自由視点画

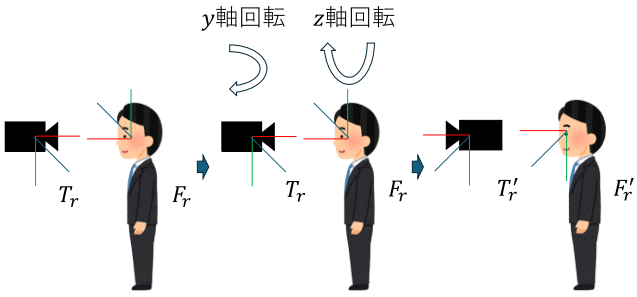

図 2 横から見たラメラ向き補正手法

† 東京工科大学 Tokyo University of Technology

像の生成において，カメラの位置と向きの情報から光線の色情報と密度情報を予測することで，精度の高い自由視点画像の生成が可能である．

続いて，被写体画像の生成について述べる．NeRF にて生成された光線空間に対し，被写体の顔の正面T_{img}から視線生成を行う．顔の正面のカメラ行列T_rは顔の向きに対する行列F_rにおいて位置関係に有意な差がないものとする．したがって，本手法では顔の正面T_rと顔の位置F_rは同じ位置であると仮定する．続いて，被写体視点のカメラの向きについて述べる．カメラの向きの補正について図示したものを図 2 に示す．カメラの向き R は以下の式で表すことができる．

$$R = T_r' F_r' T_r'^{-1} \quad (1)$$

被写体の正面の画像T_{img}に対し，深層学習型の顔姿勢推定の手法である 6DRepNet(6D Rotation Representation For Unconstrained Head Pose Estimation)[3]を用いた顔姿勢推定を行う．この時得られた回転行列をF_rとする．T_rとF_rは回転ベクトルの向きが異なるため一致させる必要がある．そのため，F_rのz軸にπラジアン回転させる．回転させた行列をF_r'とする．さらに示している角度も一致させる必要があるため，T_rのy軸にπラジアン回転させる．回転させた行列をT_r'とする．このとき，行列F_r'に対して行列T_r'による(1)の共役変換を行うことによって顔位置のカメラ行列を求めることができる．取得した顔位置のカメラ行列を用いて再度ニューラル場から被写体視点画像を生成する．

以上の通り，提案手法ではまず，NeRF を用いて生成されたニューラル場に対し，被写体の正面の視点を切り出す．続いて，切り出した被写体の正面の視点から顔姿勢推定技術を用いて顔の向きを取得する．取得した顔の向きに対して被写体の正面のカメラ行列による共役変換を行うことで被写体視点のカメラ行列が取得できる．導出した被写体の視点のカメラ行列に対して，ニューラル場から任意視点画像を生成することで被写体視点画像の生成が可能となる．

4. 実験

4.1 実験環境

本実験では，メモリを削減する観点から切り出した画像にバイキュービック法を用いて 960x540 に縮小し，実験を行った．

続いて，深層学習を用いた自由視点画像の生成について述べる．NeRF には入力データとしてカメラの位置情報を入力する必要がある．そのため本実験では，COLMAP[4]を用いてカメラの位置情報の推定を行った．深層学習を用いた自由視点画像の生成は InstantNGP[5]を用いて実施し，Iteration 数を 550000 回，ハッシュマップのサイズは 16，最適化関数は Adam，学習率は1.0×10^{-2}とした．

また，顔姿勢推定は，6DRepNet を用いて実施した．AFLW2000-3D データセット[6]を用いて学習を行った．Epoch 数を 80 回，Batch Size を 80，学習率は1.0×10^{-4}とした．

実験環境は Nvidia 6000Ada(48GB)を搭載した Linux 環境を利用し，NeRF の学習と視点生成，顔姿勢推定を行った．

4.2 生成結果

実際に本提案手法を用いて生成した被写体視点の画像を図 3 に示す．図 3 の左図は被写体が実際に見ている視点である．そして，右図には，本手法を用いて生成した被写体視点画像を示している．

(a) 実際の視点　　(b)生成した視点

図 3 実験結果

実験結果から，被写体の視点で認識しているぬいぐるみが生成結果にも含まれていることが明らかである．これらの結果から本提案手法は被写体の視点を生成できていると考えられる．しかしながら，被写体の正面のカメラ位置と顔の位置が厳密には異なるため生成した視点と実際の視点の乖離が生じてしまったと考えられる．

5. 結論

本稿では，自由視点画像の生成技術と顔姿勢推定の技術を用いて被写体視点画像生成を行った．本技術を利用すれば対象者の視点を外部の人が見ることができるようになり，防犯抑止に加え，講義中における生徒の集中度の確認，会議中でのいねむりや必要のないタブレット端末操作の特定，競技中のスポーツ選手の視線から入手する情報の特定など幅広い分野で利用が可能である．

本技術の技術的課題は以下の通りである．

まず，本システムでは被写体の顔の位置について考慮されていない．そのため，実際の被写体の顔の視線と生成された視点にわずかながら乖離が生じてしまう．

また，顔姿勢推定と自由視点画像生成でそれぞれ深層学習を用いているため，学習時間に多くの時間を要してしまう．

さらに，被写体の正面の切り出しを手動で行っているため誤差や手間が発生してしまう．そのため，より実用的なシステムにするために UI 面のさらなる改良が必要であることである．

今後はこれらの課題に取り組む予定である．

参考文献

[1] Pengfei Jia, Yun Tie, Lin Qi, et al, “PV-YOLO: An Object Detection Model for Panoramic Video based on YOLOv4”, 2022 Asia Conference on Algorithms, Computing and Machine Learning (CACML)

[2] Ben Mildenhall and Pratul P. Srinivasan and Matthew Tancik et al, “NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis”, ECCV(European Conference on Computer Vision)2020

[3] Thorsten Hempel, Ahmed A. Abdelrahman, Ayoub Al-Hamadi “6D Rotation Representation For Unconstrained Head Pose Estimation”, 10.1109/ICIP46576.2022.9897219, (2022 IEEE International Conference on Image Processing (ICIP))

[4] Schonberger, Johannes Lutz and Frahm, Jan-Michael, “Structure-from-Motion Revisited”, CVPR (Conference on Computer Vision and Pattern Recognition)2016

[5] Thomas Muller and Alex Evans and Christoph Schied et al “Instant Neural Graphics Primitives with a Multiresolution Hash Encoding”, ACM 2022

[6] Face Alignment Across Large Poses: A 3D Solution, http://www.cbsr.ia.ac.cn/users/xiangyuzhu/projects/3DDFA/main.htm