Vol.2020-HCI-189 No.19
2020/9/8

デジタルツインを用いた映像解析AI導入支援技術の研究

池田 佳弘[1,a)]　史 旭[1]　三上 啓太[1]　江田 毅晴[1]　坂本 啓[1]　榎本 昇平[1]

概要：近年，AIを用いた映像解析が注目を集めており，マーケティングなど人を対象としたAIによる映像解析の需要が高まっている．このような映像解析AIを現場へ導入する場合，カメラの配置や，照明の明るさなどの様々な要因により，AIの精度が期待する精度に到達しないことがある．この精度の低下を防ぐためには，解析対象の現場に合わせてカメラの位置や向き，画角，露出，使用するAIのモデル，推論時におけるAIの閾値などのパラメータを調整する必要がある．調整するパラメータは解析対象とする現場の状況や，カメラ機材などにより異なるため，マニュアル化が難しく，作業者の経験と勘に依存して行われる．しかし，作業熟練者の稼働費はカメラ1台につき数十万円程度要し，コストが大きい．加えて，調整時の現場状況を運用時と同じにすることは困難であり，精度低下を完全に防ぐことは難しい．そこで本稿では，解析対象とする現場の様々な状況を仮想空間上に再現した環境（＝デジタルツイン環境）を構築し，その環境内でパラメータの調整を自動で行なうシステムを提案する．また，予備実験により，デジタルツイン環境をAIの精度検証環境に利用することが有効であることを示す．提案するシステムにより，パラメータの設定を誰でも簡易に行なうことを可能とし，映像解析AI現場導入時の障壁を低減する．

キーワード：映像解析AI，導入支援，デジタルツイン，稼働削減

1. はじめに

近年，AI（人工知能）を活用した映像解析の市場規模は拡大傾向にある．ミック経済研究所の調査によると，2018年度に53億円であった国内におけるAIを活用した映像解析（以下，映像解析AI）の市場規模は，図1に示すように2023年度までには1500億円になると予測されている[1]．このような状況の中で，セキュリティやマーケティングなどの人を対象とする映像解析の需要も高まっており，映像解析AIの社会実装は今後一気に進展することが予想される．

映像解析AIをビジネスの現場に導入する場合，AIの精度は高いことが期待されるが，実際には期待する精度よりも低下してしまうことがほとんどである．その理由としては，カメラと解析対象までの距離や，照明の明るさ，オクルージョン等の現場の様々な要因が，AIの精度に影響を与えるためである．従って，映像解析AIの現場導入時においては，カメラを設置する位置や向き，画角，露出，使用するAIのモデル，推論時におけるAIの閾値（例えば，人物照合の例においてはスコアが何%以上であれば同一人物と判定するかなどに利用される）など，複数のパラメータを調整して組み合わせ，精度の低下を防ぐための作業（以下，調整作業）を行う必要がある．

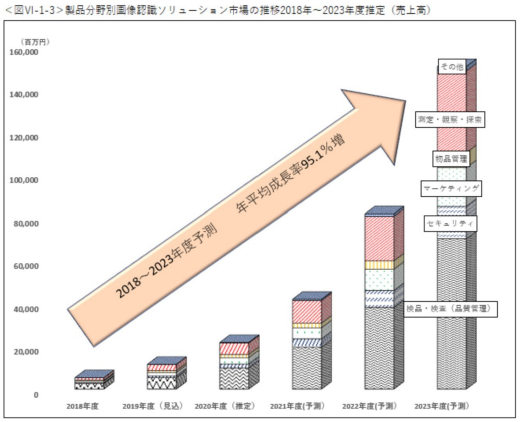

図 1　映像解析AI市場規模の予測[1]

しかし，調整作業には2つの課題がある．(1) 調整作業

[1] NTTソフトウェアイノベーションセンタ第二推進プロジェクト
3–9–11 Midori-cho, Musashino-shi, Tokyo, 180-8585 Japan
[a)] yoshihiro.ikeda.ap@hco.ntt.co.jp

は導入先の現場状況によって調整するパラメータが異なるため，マニュアル化が難しく，時間単価が高い作業熟練者の勘と経験に依存して行われる．結果として，カメラ 1 台当たりの調整作業にかかる熟練者の稼働費は，市場価格で数十万円程度要し，コストが大きい．また，パラメータを調整するごとにエキストラ等を撮影して AI の精度を確認する必要があり，多くの時間を要する．(2) AI の精度検証を行なう現場の状況が限られており，調整作業を行なったとしても精度低下を完全には防ぐことは難しい．これは例えば，調整作業を店舗に客のいない深夜に行なうといった場合に，調整時（夜）と運用時（昼）の照明状況が異なること，エキストラの服装や性別，人数などの現場状況のバリエーションを運用時と同じにすることが難しいことなどが挙げられる．

これに対し，近年では仮想空間上に実空間を再現し，仮想空間上でシミュレーションした結果を実空間へ反映するデジタルツイン [2] というコンセプトが着目されている．仮想空間上には，現場の様々な状況を実空間より低コストで再現可能になるほか，各種パラメータの調整にかかる膨大な試行錯誤の実行も自動化できると考えられる．

そこで本研究では，解析対象とする現場を仮想空間上に構築し，人の 3D モデルを配置して現場の様々な状況を再現した環境（=デジタルツイン環境）を構築し，人の 3D モデルに対して，映像解析 AI の精度を高める調整作業を自動で行うシステムを提案する．このシステムの出力結果に従うことで，カメラ位置等のパラメータの設定を，誰でも簡易に行なうことを可能とし，熟練者の作業にかかるコストが大きいことと，調整時の現場状況を運用時と同じにすることが難しいことの 2 つの課題を解決する．

本稿の以降の構成を記す．2 章では関連研究と本研究の位置づけについて述べる．3 章では提案するシステムについて述べる．4 章ではデジタルツイン環境上で映像解析 AI の精度検証が可能か，実空間とデジタルツイン環境で映像解析 AI の出力が異なるかについて検証した予備実験について述べ，5 章ではまとめと今後の検討について述べる．

2. 関連研究

2.1 仮想空間上で AI の学習データを作成する研究

実空間で大規模なラベル付きデータセットを作成することは容易ではないため，近年では仮想空間上で作成したデータを学習に利用することに注目が集まっている．例えば，Veeravasarapu ら [3] は，CG データの輝度値などを GAN を用いて実写と近づくように調整する手法を提案した．また，Dai ら [4] は AI による 3D シーンの理解に向けて，セマンティックセグメンテーションでインスタンスレベルにアノテーション付けされた 3D モデルから構成される 1500 ものシーンを，250 万もの視点から撮影した大規模なデータセットを作成した．しかし，以上の研究の目的はあくまでも学習データの作成である．本研究では，仮想空間を学習済み AI モデルの精度検証に利用するという点で，既存研究と目的が異なる．

2.2 仮想環境上におけるカメラの最適配置

仮想空間上において，カメラの最適な位置や向き，ズームの値を探索する研究も行われている．Bisagno ら [5] は粒子群最適化手法を用いて，Unity 上に構築した仮想の屋内環境において，照明環境や床，壁などの障害物を考慮し，同一の仮想 PTZ カメラ複数台の視野領域を最大化するカメラ配置を求める手法を提案した．しかし，この研究の目的は，利用可能な最小限のカメラ台数で監視するエリアにおいて死角をなくすための最適配置を探索することであり，対象は複数台のカメラである．また，映像解析 AI の精度を高めることについては考慮していない．本研究では映像解析 AI を現場に導入する場合を対象とし，個々のカメラについて人を対象とした映像解析 AI の精度を高めるための各種パラメータを探索するという点で，既存研究とは異なる．

3. 提案するシステム

本研究が提案するシステムの詳細と，システムを構築するための要素技術について述べる．

3.1 提案するシステムの詳細

本研究が提案するシステムは，(1) 解析対象とする現場を仮想空間上に再現し，人の 3D モデルを配置したデジタルツイン環境を構築すること，(2) 構築したデジタルツイン環境内で，人の 3D モデルに対する映像解析 AI の精度を高めるための様々なパラメータの最適値を自動で探索することの 2 つのステップにより実現する．

提案するシステムのイメージを図 2 に示す．このシステムの入力は，カメラや距離センサなどから取得した画像や深度情報，点群といったデジタル情報である．このデジタル情報を用いて，現場で想定されるシチュエーションに合わせて，人の 3D モデルが配置されたデジタルツイン環境を構築する．そして，使用する AI のモデルと，期待する精度や探索の試行回数などの探索の終了条件を設定する．デジタルツイン環境上ではパラメータを探索しつつ，各パラメータの設定でデジタルツイン環境内を撮影した画像に対し，AI の精度を算出する．この探索は終了条件を満たすまで行い，最終的に AI の精度が最も高い状態におけるパラメータ（位置，向き，画角，AI モデル，AI モデルの閾値）を出力する．

本システムを導入することにより，パラメータの調整をデジタルツイン環境内で自動化することで，作業熟練者でなくても簡易に AI の精度が高くなるように各パラメータを設定することが可能になる．これにより，カメラ 1 台当

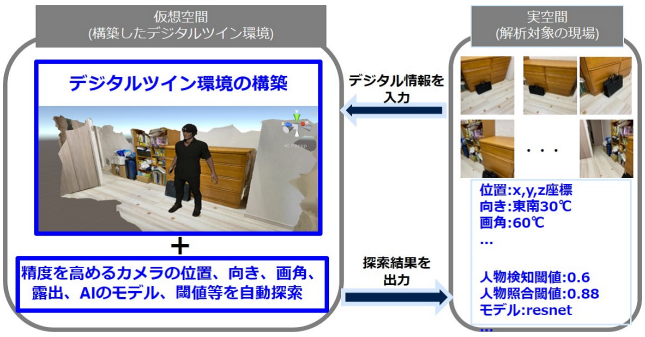

図 2　提案するシステムのイメージ図

図 3　構築したデジタルツイン環境の外観

たりにかかる作業熟練者の稼働費をなくし，調整作業に関わるコストを抑える．さらに，デジタルツイン環境上で様々な現場状況を再現し，現場状況のバリエーションを増やすことで，AI の精度低下を抑えることが可能になると考えられる．

3.2　仮想空間上に実空間を再現する技術

センサによって取得した情報から，実空間上の 3 次元形状を仮想空間上に再構築する既存研究は様々なものが存在する．これらの手法は大きく 2 つに分けられる．まず，あらかじめ取得済みの未整列な画像群からカメラの位置と向きを推定し，その情報をもとに撮影対象の密な 3 次元形状を復元する手法である．この手法は Photogrammetry と呼ばれており，Photogrammetry を行なうためのソフトウェアやライブラリとしては，3DF Zephyr [6] などの様々なものが存在する．次に，画像や深度情報の取得と位置・向き推定，3 次元形状の復元をリアルタイムに行う手法である．この手法は SLAM と呼ばれており，カメラから得られる情報のみからカメラの位置と向きを推定する Visual SLAM [7,8] や，深度センサを用いてカメラの位置，向きを推定する手法 [9] などが存在する．本研究ではこれらの手法を活用し，解析対象の現場を仮想空間上に再現する．

3.3　パラメータを探索するための技術

調整作業を自動で行うためのアプローチとしては，すべてのパラメータの組み合わせを試し，その中で最も良いスコアを出す組み合わせを選択するグリッドサーチと呼ばれる方法や，パラメータの組み合わせを最適化問題として解く，粒子群最適化手法 [10] などが考えられる．

4.　予備実験

4.1　実験目的

本研究で提案するシステムでは，構築したデジタルツイン環境のパラメータ探索結果が実空間に適用可能なことを要件として達成する必要がある．しかし，実空間とデジタルツイン環境にはギャップがあるため，デジタルツイン環境と実空間でパラメータを揃えたとしても，AI の出力が異なる可能性があり，デジタルツイン環境上で求めたパラメータの設定が実空間に適用できるかは明らかではない．

そこで本稿では，デジタルツイン環境では AI の精度検証がそもそも可能か，デジタルツイン環境と実空間において，それぞれ AI の出力が近づくのかを実験により検証した．

4.2　実験内容

本実験では，実空間とデジタルツイン環境においてカメラの内部パラメータ，位置，向きの設定を揃えてカバンを撮影し，撮影した 2 種類の画像に対して，物体検出を行なう映像解析 AI を使用した際の出力を比較した．

まず，解析対象の現場を再現したデジタルツイン環境の構築に向けて，事前準備として，現場の床にカメラを設置する位置の目印を設置した．この目印はカバンを中心に，30° の方向に 60cm 離れた位置に用意した．そして iPhone 11 Pro を用いて，カバンを設置済みの室内を各画像が 6 割程度重なるように環境全体を隙間なく撮影した．さらに，撮影した計 301 枚の画像から，3DF Zephyr [6] を用いてデジタルツイン環境を構築した．なお，構築したデジタルツイン環境は Unity 上にインポートし，デジタルツイン環境の縮尺と現場の縮尺を手動で揃えた．図 3 に構築したデジタルツイン環境の様子を示す．

次に，現場におけるカメラと，Unity 上における仮想のカメラの内部パラメータ，位置，向きが同じになるよう設定した．まず，カメラの内部パラメータを合わせるため，現場におけるカメラの内部パラメータをカメラキャリブレーションにより求め，仮想のカメラへ反映した．また，位置を合わせるために，現場のカメラと仮想のカメラを，目印の垂直方向に，高さ 36cm のところに設置した．さらに，各カメラの角度を，床と垂直の軸に沿って 30° でカバン方向へ向けた．

最後に，現場で撮影した画像と，デジタルツイン環境で撮影した画像の計 2 種類の画像に対し，物体検出を行なう映像解析 AI である YOLO v3 [11] を使用した際の，Bounding Box の分類ラベルとクラス確率の比較を行なった．

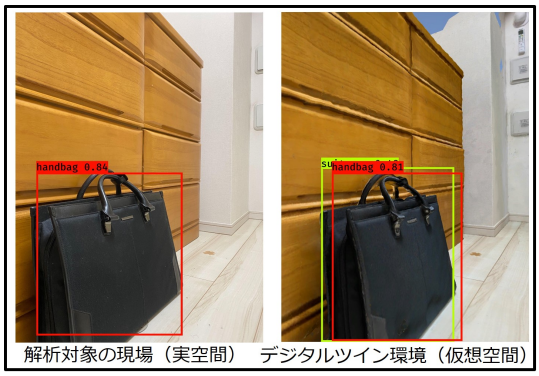

図 4　出力の可視化結果

4.3　実験結果

図 4 に出力の可視化結果を提示する．現場を撮影した画像に対する結果を見ると，1 個の Bounding Box がカバンを囲っており，分類ラベルは handbag，クラス確率は 0.84 であった．一方，デジタルツイン環境を撮影した画像においては，カバンを 2 個の Bounding Box が囲っており，分類ラベルはそれぞれ handbag，suitcase であり，クラス確率は 0.81，0.48 であった．

4.4　考察

どちらの画像においても分類ラベルが handbag の Bounding Box が存在し，かつその Bounding Box におけるクラス確率の差は 0.03 という極めて小さい値であったことから，デジタルツイン環境上で AI の精度検証は可能であり，その出力は実空間における出力と近くなる可能性があることが考えられる．ただし，今回の実験では比較に使用した画像は 1 つの位置と向きから撮影した画像のみで検証した結果であり，位置や向きが異なった場合にも同様の結果が得られるかは明らかでない．そこで今後は，撮影するカメラ位置のバリエーションを増やして検証を行う必要がある．

また，今回の実験では，解析対象とするカバンを現場環境に含めてデジタルツイン環境として構築した．しかし，提案システムの解析対象は人物であり，現場環境構築時，現場で常に移動する人間を埋め込みながらデジタルツイン環境として構築することは不可能であるため，人の 3D モデルを現場環境と独立に用意する必要があると考える．ここで，別環境で作成した人の 3D モデルは現場環境の照明状況と異なる可能性があるため，現場環境の照明に合わせて人の 3D モデルに対する輝度などの調整が必要となることも考えられる．

加えて，今回の実験ではデジタルツイン環境の構築に向けて，各画像が 6 割重複するよう 300 枚程度の画像を撮影を行なった．実際の映像解析 AI 導入先は，予備実験で構築した空間より遥かに広いことが想定されるため，大量の現場写真を用意する必要があることが想定される．そのため，今後は少ない画像数で簡易にデジタルツイン環境を構築する方法についても検討する必要があると考える．

5.　まとめ

本研究では，映像解析 AI の現場導入時に行う調整作業において，熟練者に掛かる稼働費が大きことと，現場作業の制限により精度改善が難しいことというの 2 つの課題を，デジタルツインを利用することで解決するシステムを提案した．また，予備実験を通じて，デジタルツイン環境における映像解析 AI の出力と，実空間上における AI の出力は近くなる可能性を示した．今後は，構築した現場の仮想空間に，別で用意した人の 3D モデルを，様々な現場状況に合わせて配置し，様々なカメラ位置及び向きでデジタルツイン環境上の AI 出力結果が実空間上の AI 出力結果と同じ傾向であるかを検証する予定である．

参考文献

[1] MIC research institute ltd. online. (2019). AI（ディープラーニング）活用の画像認識ソリューション市場の現状と展望【2019 年度版】. Retrieved from https://mic-r.co.jp/mr/01760/ (最終アクセス日 2020/5/31)
[2] Grieves, M. (2015). Digital Twin: Manufacturing Excellence through Virtual Factory Replication.
[3] Veeravasarapu, V.S., Rothkopf, C.A., & Ramesh, V. (2017). Adversarially Tuned Scene Generation. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 6441-6449.
[4] Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T.A., & Nießner, M. (2017). ScanNet: Richly-Annotated 3D Reconstructions of Indoor Scenes. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2432-2443.
[5] Bisagno, N., & Iacovlev, C. (2019). Camera network optimization: maximize coverage in a 3D virtual environment. Proceedings of the 13th International Conference on Distributed Smart Cameras.
[6] 3D Flow. online. (2017). 写真計測用ソフトウェア 3DF ゼファー (3DF Zephyr). Retrieved from http://www.opt-techno.com/opt-zephyr/ (最終アクセス日 2020/5/31)
[7] Mur-Artal, R., & Tardós, J.D. (2017). ORB-SLAM2: An Open-Source SLAM System for Monocular, Stereo, and RGB-D Cameras. IEEE Transactions on Robotics, 33, 1255-1262.
[8] Rückert, D., Innmann, M., & Stamminger, M. (2019). FragmentFusion: A Light-Weight SLAM Pipeline for Dense Reconstruction. 2019 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct), 342-347.
[9] Dai, A., Nießner, M., Zollhöfer, M., Izadi, S., & Theobalt, C. (2017). BundleFusion: Real-Time Globally Consistent 3D Reconstruction Using On-the-Fly Surface Reintegration. ACM Trans. Graph., 36, 24:1-24:18.
[10] Kennedy, J., & Eberhart, R. (1995). Particle swarm optimization. Proceedings of ICNN'95 - International Conference on Neural Networks, 4, 1942-1948 vol.4.
[11] Redmon, J., & Farhadi, A. (2018). YOLOv3: An Incremental Improvement. ArXiv, abs/1804.02767.