一般社団法人 電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report

注釈情報を３次元的に AR 提示可能な多視点映像閲覧方式

志田 全弘† 亀田 能成† 石塚 祐香‡ 辻 愛里†
榎本 拓哉‡ 山本 淳一‡ 鈴木 健嗣† 北原 格†
†筑波大学 〒305-8571 茨城県つくば市天王台 1 丁目 1-1
‡慶應義塾大学 〒108-8345 東京都港区三田 2 丁目 15-45
E-mail: †s1520786@u.tsukuba.ac.jp

あらまし

Bullet-Time 映像上に３次元的な注釈情報を AR 提示する手法について述べる．近年，撮影空間を様々な角度から観察可能な多視点映像を用いた映像メディアの要求が高まっている．多視点映像の閲覧方式の中でも，Bullet-Time 方式は映像品質が高く，撮影空間内のオブジェクトの位置関係の把握が容易であるという特長を有する．一方で，Bullet-Time 方式は撮影シーンの３次元情報を陽に復元しないため，AR 提示における幾何学的整合の再現が困難であるという問題がある．本研究では，多視点映像から撮影空間の大まかな３次元情報を推定し，映像理解を支援する注釈情報を３次元的に提示する機能を Bullet-Time 方式に追加することを目的とする．提案手法では，画像特徴の対応付けから撮影空間の疎な３次元情報やカメラポーズを推定する Structure from Motion と密な３次元情報を推定する Multi View Stereo 法を用いて３次元情報を推定する．推定した３次元情報に基づき映像内の人物が注目している物体の推定や隠れの再現を行う．

キーワード バレットタイム，多視点映像，注釈情報，拡張現実，幾何学的整合性，隠れの再現

1. はじめに

本稿では多視点映像の閲覧方式の一つである Bullet-Time 方式[1]に注釈情報を３次元的に重畳する手法について述べる．視聴者の嗜好の多様性に応えるべく，撮影空間を様々な角度から観察可能な多視点映像メディアへの期待が高まっている．多視点映像を用いた映像メディアは，多面的な観察が可能であるため，「シーンの理解促進」や「新たな気づきの獲得」といった効果が期待される．多視点映像の閲覧方式には，監視カメラのモニタリングシステムに代表されるマルチウィンドウ方式や，自由視点映像技術[2]に代表される視点選択方式がある．マルチウィンドウ方式は，人が同時に観測可能なカメラ台数に限界があるため，不審者検知のように観察対象が少数に限定される用途以外には適さない．視点選択方式は３次元再構成処理を介すため画質が劣化しやすいという問題がある[3]．

Bullet-Time 方式は撮影対象の周囲に配置した多視点カメラで撮影した映像をカメラ配置に従って連続的に切り替えながら提示する映像技術である．撮影映像をほぼそのまま提示するため画質劣化が少なく，閲覧者は撮影空間を３次元的に移動するような視覚効果を得る．その一方，３次元的な情報を陽に持たないため撮影空間の３次元情報と整合した AR(Augmented Reality)提示が困難であるという問題を持つ．AR は実写映像に仮想情報を重畳することにより視覚情報の増強を実現する技術である．本研究では，多視点映像から撮影空間の大まかな３次元情報を推定することで幾何学的整合性の問題を解決し，３次元的な AR 提示が可能な Bullet-Time 方式の実現を目指す．

撮影対象として，複数人によるコミュニケーションに着目する．「シーンの理解促進」や「新たな気づきの獲得」を促す注釈情報を AR 提示することで，映像を用いた状況分析の支援を行う．状況分析では，人物の注目方向や表情・体の強張りなどが表面化した人物の心理状態に注目が集まる[4][5]．また，コミュニケーション時の各人物の立場や役割も重要な情報となる．我々は，それらの撮影対象の情報を的確に伝達するためには，以下３点に考慮した映像提示が有効であると考えている．

(1) シーンや状況理解の支援
(2) 注目している物体の明確化（強調）
(3) 注目対象の詳細な観察

(1)については，撮影した人物にキャプションを付与し，閲覧者が各人物の名称やシーンにおける役割などを把握することを支援する．(2)では，人物の注目している物体の強調提示や注目方向を AR 提示し，注目方向の推測を支援する．(3)では，注目箇所をデジタルズームすることで表情や体の緊張などといった詳細な観察を支援する．(1)と(2)は，AR 提示機能の追加により実現を試みる．(3)はズーム観察が可能な Bullet-Time 方式[6]を採用することで実現する．

2. 関連研究

2.1. Bullet-Time 方式の利用

Bullet-Time 方式は映画 The Matrix で使われたことで広く知られるようになり，アメリカンフットボールや

This article is a technical report without peer review, and its polished and/or extended version may be published elsewhere.
Copyright ©20●● by IEICE

野球の試合中継にも用いられている[7][8]．近年では，ミュージックビデオや CM などにも利用されるなど実用化が進んでいる [9].

2.2. 作業支援としての映像メディア

作業支援を行う現場では少数台のカメラで撮影した映像を用いて状況分析が行われている[4][5]. 本研究が対象とする複数人がコミュニケーションを行う環境でも，行動に対する反応の良し悪しや，注意力の持続時間などが，音声や映像情報を元に分析されている．しかし，少数台のカメラを用いた撮影では，人物の移動や顔の向き，人物同士の重なり合いなどにより十分な分析が難しい状況が発生しやすい．

Yonemoto ら[10]は，単眼視点の映像に注釈情報を AR 提示することで映像分析を支援している. しかし，画像特徴を用いた追跡処理によって注釈情報の提示位置を推定しているため，提示対象が平面に限定され，かつ,物体の隠れには対応できないという問題がある．

本研究では多視点映像を用いて作業内容の理解を促進する閲覧方式を提案する. さらに Bullet-Time 方式により撮影空間を多角的かつ直感的に観察可能な教示映像メディアを提案する．

2.3. 3次元情報推定技術

撮影シーンの３次元情報推定法として，Structure from Motion（SfM）と Multi View Stereo 法（MVS）を組み合わせる手法が提案されている[11][12]. SfM は多視点映像から対応点を推定し，各カメラの位置や姿勢と撮影空間の疎な３次元点群を推定する手法である．視点数の増加に伴い誤差が蓄積していく loop closure やテクスチャレスな映像では推定精度が低下するという問題がある．MVS は SfM により推定されたカメラポーズと疎な点群を利用し，密な点群の推定やメッシュの生成，テクスチャの貼り付けを行う．誤推定によるメッシュやテクスチャの欠損などの問題が存在する．どちらも画像特徴点が安定して検出可能な映像の利用を前提としているが，本研究で対象とする撮影シーンが必ずしもそのような状況にあるわけではない．そこで，テクスチャ情報を豊富に有する物体を環境内に多数配置した状態でカメラキャリブレーション用の撮影を行い，そのカメラパラメータに基づいて SfM・MVS を実行することにより, 3 次元推定精度を向上させる．

3. 撮影空間の３次元情報の推定

3.1.　３次元情報推定の安定性向上

撮影中はカメラの位置姿勢が変化しないことを前提として，カメラパラメータ推定を行う際には，撮影シーンに画像特徴の出やすい（テクスチャ情報を多く含む）オブジェクトを多数配置した状態で多視点映像を撮影し，SfM によるカメラキャリブレーションの精度を向上させる．

Bullet-Time を生成するシーンの３次元情報を推定する場合は，先に推定したカメラパラメータと画像特徴点を用いて隣接する２台のカメラ間の対応点探索を行う．この時，対応関係にある第一候補と第二候補の対応点の類似度に大きく差異があるもののみを対応付けることで，不安定な対応付けを避ける．カメラパラメータと２視点画像の対応点から対応点の３次元情報を推定する．同様の処理を n 台のカメラ全てに行うことで撮影空間の疎な点群を推定する．

3.2.　モーションキャプチャを用いた３次元計測（注釈情報の獲得）

赤外線カメラと再帰性反射材のマーカで構成される光学式モーションキャプチャを用いて被写体の３次元位置を計測する．３次元センサの座標系をセンサ座標系とする．センサ座標系の XY 軸を規定する, 図 1 に示すような L 字型にマーカ（原点 $\boldsymbol{B_0}$, X 軸 $\boldsymbol{B_1}$, Y 軸 $\boldsymbol{B_2}$）が配置された機材を用いて赤外線カメラのキャリブレーションを行う．赤外線カメラから再帰性反射材のマーカへ赤外線を照射し，その反射光を複数の赤外線カメラで撮影する．カメラキャリブレーションで求めたカメラの射影変換行列 P, P’ と２台のカメラで検出したマーカの観測位置 m, m’（２次元座標）からマーカの３次元位置 X を求める．

図 1.モーションキャプチャの基準座標系を定める L 字型のキャリブレーションツール

3.3.　３次元モデル生成時の座標系とセンサ座標系の統合

SfM で定義される３次元モデルの座標系（モデル座標系）とセンサ座標系の間の剛体変換行列を推定する．センサ座標系のキャリブレーションで用いた

３点（原点 $\boldsymbol{B_0}$, X 軸 $\boldsymbol{B_1}$, Y 軸 $\boldsymbol{B_2}$）と対応するモデル座標系の点 $\boldsymbol{A_0}$, $\boldsymbol{A_1}$, $\boldsymbol{A_2}$ の３次元座標をステレオ視で求める．対応点情報から座標系間のスケールパラメータを算出する．座標系間の回転変換行列は，各軸の正規化方向ベクトルにスケールパラメータをかけたベクトル $A'_{ox}, A'_{oy}, A'_{oz}$ によって構成される式(3)の直交行列として求まる．この回転変換行列 $\boldsymbol{H}$ と原点座標値を平行移動量とする平行移動ベクトル A_o によって，式(4)のように任意のセンサ座標系の点 Bp をモデル座標系 Ap に移動することができる．

$$\boldsymbol{H} = \begin{pmatrix} A'_{ox_1} & A'_{ox_2} & A'_{ox_3} \\ A'_{oy_1} & A'_{oy_2} & A'_{oy_3} \\ A'_{oz_1} & A'_{oz_2} & A'_{oz_3} \end{pmatrix} \tag{3}$$

$$A_p = \boldsymbol{H}B_p + A_o \tag{4}$$

4. ３次元的な注釈情報の生成と提示

4.1　３次元注釈情報の生成

複数人がコミュニケーションを行う空間において，対話している人物の名前や視線方向，注目しているものを把握することは相手を理解する上で重要な情報である．本研究では，それらの状況分析を支援する注釈情報を図 2 に示すように AR 提示する．

図 2.提案手法による注釈情報の提示

4.1.1　人物情報の AR 提示

文字を重畳する場合，背景画像の色味によって見えやすさが変化する．そこで図 2 に示したように背景と前景（文字）の色に明暗差を付けた名札のような画像を用いる．これをプレート画像と呼ぶ．３次元的にプレート画像を重畳する場合，Bullet-Time 方式では視点が切り替わることでプレート画像も回転してしまい，注釈情報の視認性が低下する．そこで，CG モデリングで広く利用されるビルボード法を適用し，全ての視点においてカメラの光軸方向と正対した提示を実現する．

4.1.2　視線方向の AR 提示

３次元的な視線方向の取得には二つのアプローチがある．一つが人物の瞳孔検出である[13]．視線ベクトルを直接的に求めることができるが，カメラと被写体に距離がある Bullet-Time 映像では瞳孔周辺の解像度が低く適用が困難である．そこで本研究では，モーションキャプチャにより人物の頭部位置と方向を取得し，この方向を人物の視線方向として扱う．視線ベクトルの始点と方向から３次元矢印を提示する．

4.1.3　注目物体の強調提示

注目物体の色を変化させることで強調提示を実現する処理では，視認性の結果が背景色に依存してしまう.そこで本研究では背景画像の色をそのまま利用し，その代わり図 3 に示すように，背景画像の各色チャンネル（RGB）の値を n 倍（n＞1.0）することにより，注目物体の周囲の明度を高くすることで強調提示を実現する．

図 3.注目物体の強調提示

4.2　注目物体の推定

視線方向と撮影空間の３次元情報を手掛かりに注目物体を推定する．視線方向を延長し，３次元点群との交点を推定する．本手法では，図 4 に示すように，延長する直線から最も近い３次元点群までの距離が閾値 k 以下となるとき，３次元点群と直線が衝突したものとみなし，注目物体を構成する点群を推定する．

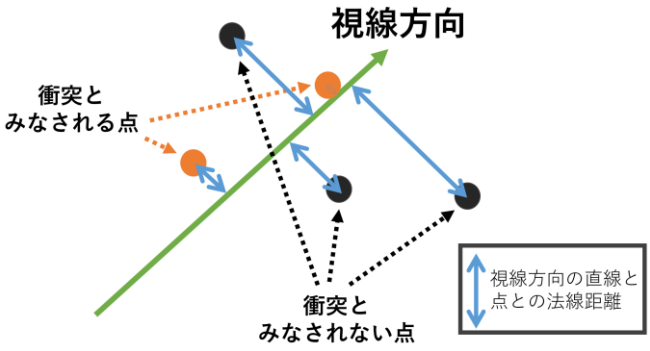

図 4.視線方向上の３次元点群の検出

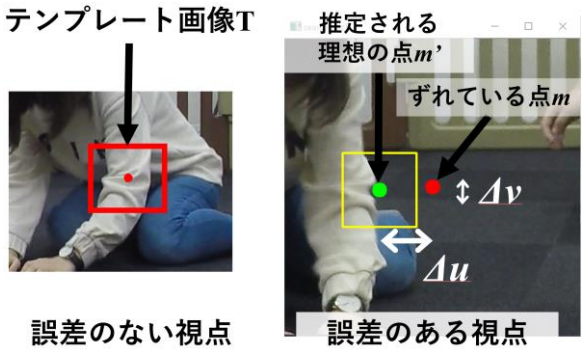

図 5.テンプレートマッチングによる
投影位置のズレ($\Delta u, \Delta v$)の推定

4.3　注釈情報の提示位置の補正

SfM で推定したカメラパラメータには，loop closure などによって生じる誤差が含まれている．loop closure はカメラ台数の増加に伴い誤差が蓄積し，実際のカメラ位置と推定されたカメラ位置に誤差が生じる問題である．その結果，画像上で実際に観測される２次元座標からズレた位置に３次元点が投影されてしまう．

この問題を解決するため隣接するカメラの見え方の類似性を利用する．Bullet-Time 映像提示を想定して撮影した多視点映像では，隣接するカメラ間隔が比較的狭い．そこで図 5 に示すように隣接画像間でテンプレートマッチングを行い，注釈情報の提示位置を補正する並進移動行列 $\boldsymbol{D}$ を推定する．並進移動行列 $\boldsymbol{D}$ は式(5)によって表される．

$$\boldsymbol{D} = \begin{pmatrix} 1 & 0 & \Delta u \\ 0 & 1 & \Delta v \\ 0 & 0 & 1 \end{pmatrix} \qquad (5)$$

ここで Δu 及び Δv はある３次元点の射影変換後の２次元座標 m と理想的な観測位置の２次元座標 m' との差分である．３次元点はテンプレートマッチングの精度向上のため輝度勾配の激しい領域中の点を用いる．

注釈情報を提示する３次元点を射影変換し，その同次座標に並進移動行列 $\boldsymbol{D}$ をかけることで注釈情報の提示位置を補正する．

4.4　カメラ位置に応じた隠れの表現

提示するカメラと被写体の位置関係によって注釈情報の手前に実物体が存在する場合があるため，図 6 に示すように透明度の変化によって物体との隠蔽処理を表現する．注釈情報（仮想物体）と実物体の前後関係を判定するために，カメラ位置と撮影空間の３次元点群を用いる．カメラ位置と注釈情報の３次元位置との差分から注釈情報へのベクトルを計算し，そのベクトルと３次元点群を 4.2 節の手法に適用しベクトル上の３次元点群を検出する．検出された３次元点とカメラとの距離，注釈情報の３次元位置とカメラとの距離から前後関係を判定する．

注釈情報の CG の色を $C(r,g,b)$ とする．実物体よりも後方にある（隠れている）と判定された CG は，式(6)を用いて半透明化処理を施す．ここで背景画像の色を C_{img}，提示する CG の色を $C_{bg}(r_{bg}, g_{bg}, b_{bg})$ とする．

$$C_{bg} = kC + (1-k)C_{img} \qquad (6)$$

図 6.半透明化による CG の隠蔽処理

5. 評価実験

本章では，複数人がコミュニケーションを行う環境を撮影した映像データを用いて，提案手法の実用性を検証する．また，提案手法により生成される映像の作業支援に対する有効性を検証する．

5.1　撮影実験

療育現場において複数人がコミュニケーションを行う様子を撮影し，得られた多視点画像に注釈情報を付加する．カメラパラメータの推定や３次元モデル生成のため,オープンソースの Visual SfM[14][15]と Open MVS を用いる.使用する計算機は, OS: Windows 10 Pro, CPU：Intel Core i7-6700K 4.00GHZ，GPU：Intel HD Graphics 530，メモリ：32GB[RAM]を搭載した 64bit ベースのデスクトップ PC を用いる．GoPro 社の GoPro HERO4 ブラック 20 台を撮影シーンを取り囲むように設置し,3840[pixel]×2160[pixel]の解像度の映像を毎秒 30 フレーム撮影した．モーションキャプチャは OptiTrack 社の Flex3 を使用し,天井の四隅に２台ずつ,計８台設置した赤外線カメラで計測を行う．人物の顔の向きの推定には Tsuji ら[16]の手法を用いる.

5.2　提案手法の作業支援に対する有効性

提案法式の有効性を検証するため，３種類の実験を行う．実験１では視点選択の重要性を，実験２では視点数の増加による映像理解への影響を，実験３では注釈情報の提示による映像解析への有効性を検証する.

実験１，２は普段から映像解析を行っているセラピスト１名と映像解析に不慣れな学生 13 名を対象に行った．実験３はセラピスト１名を対象に行った.

実験１，２では実際の状況分析において注目される,コミュニケーションを行う人物の表情や目線が，話し手と受け手の両者とも十分に確認できるか，５段階のリッカート尺度を用いて検証する．実験１では，被験者に８視点分の映像を１視点につき一分間ずつ提示し,各視点の映像に対する回答を得た．実験２では，五つのシーンに対し単眼視点の画像と Bullet-Time 方式に

よる多視点画像を提示し，各シーンに対する回答を得た.

5.2.1　視点選択の重要性（実験１）

実験１の回答結果を表１に示す．数値が高いことは映像理解のし易さを示す．表１より視点３が最も映像理解がし辛く，視点 1,6,7 がし易いことが分かる．視点３は隠れが発生し，後方の情報が確認できない．一方で視点６は被写体を真横から観察できるため，作業内容や人物の表情を確認可能である．以上より，映像理解のためにはどの視点で閲覧するかが重要であるといえる.

表 1．実験１の主観評価

	平均	標準偏差
視点１	3.36	1.17
視点２	3.14	0.91
視点３	1.00	0
視点４	1.64	0.61
視点５	2.79	0.77
視点６	3.64	1.04
視点７	3.57	1.29
視点８	2.64	1.17

5.2.2　視点数の増加による映像理解への影響（実験２）

実験２の回答結果を表２に示す．数値が高い程多視点映像の方が，低い程単眼視点映像映像の方が理解し易いことを示す．表２よりどのシーンの画像に対しても，単眼視点より多視点（Bullet-Time 方式）での閲覧が映像理解しやすいという結果が得られた．セラピストから「映像解析時には複数のセラピストが同じ映像について分析結果を話し合うが，多視点で閲覧することでその一致率が高まり，分析結果の信頼性の向上に繋がる」という意見を得られた．また,「被写体を回り込んで閲覧することで被写体の姿勢を観察することができ，体の緊張が確認できるなど，従来の少数視点による映像解析では得られない情報を得ることができる」という意見も得られた.

表 2．実験２の主観評価

	平均	標準偏差
シーン１	4.36	0.71
シーン２	4.86	0.35
シーン３	4.21	0.77
シーン４	4.64	0.48
シーン５	4.29	0.80

5.2.3　注釈情報の提示による映像解析に対する有効性（実験３）

実験３ではセラピストが五つのシーンに対し多視点映像解析を行い，注釈情報の提示の有無により，映像解析の時間に変化が生じるかを検証する．被写体の注目方向や注目物体を判断するまでに要した時間を図 7 のグラフに示す．グラフから注釈情報の提示により

映像解析に要する時間が減少していることが分かる．減少した時間の平均値は5.2秒であり，標準偏差は4.02であった．セラピストからは，注釈情報があることで解析が早くなること，作業支援に不慣れな人に映像説明する際に話しやすいなどの意見が得られた．

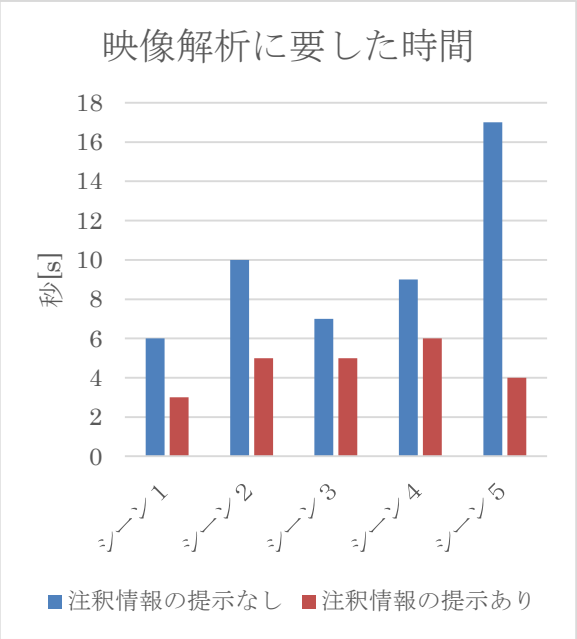

図 7.映像解析に要したシーン毎の時間

6. お　わ　り　に

Bullet-Time 方式に注釈情報の３次元的な AR 提示を行う手法を提案した．多視点映像から撮影空間の大まかな３次元情報を推定し，３次元的な注釈情報の提示を実現した．その際，隣接画像を用い画像間での対応情報を増やすことで，３次元情報の推定精度を向上させた．注釈情報の提示時には，推定した３次元情報をもとに映像内の人物の注目物体の推定や隠れの再現を行った．評価実験では複数人がコミュニケーションを行う空間を撮影した映像に本手法が適用できること，本手法により映像理解が向上されることを確認した．今後の課題として，より豊富な種類の注釈情報の提示や，その CG モデルを簡易的に導入する枠組みの開発が挙げられる．

謝　辞

本研究の一部は，科学技術振興機構(JST)の戦略的創造研究推進事業(CREST)における研究領域「人間と調和した創造的協働を実現する知的情報処理システムの構築」の研究課題「ソーシャル・イメージング：創造的活動促進と社会性形成支援」により行ったものである．

文　献

[1] 坂本竜基，陳鼎，“解像度を維持しつつ自然な Bullet Time を実現する射影変換の検討”，情報処理学会 インタラクション 2014，pp.65-70，2014

[2] T Kanade, P.W Rander, P.J. Narayanan, “Virtualized Reality: Constructing Virtual Worlds from Real Scenes”, IEEE Trans Multimedia, Vol.4, No.1, pp.34-47, 1997

[3] 秦哲也，北原格，大田友一，“自由視点映像提示における人物ビルボード方式と３次元復元方式との比較”，電子情報通信学会 2009 年総合大会講演論文集，p.189, 2009

[4] 山本淳一，大森貴秀，皆川泰代，“自閉性障害幼児の家庭訪問型発達支援モデルの構築と包括的評価”，科学研究費補助金研究成果報告書 (2013), 2014

[5] Y Ishizuka, J Yamamoto, “Contingent imitation increases verbal interaction in children with autism spectrum disorders”, Autism: International Journal of Research and Practice, 2016

[6] N Akechi, I Kitahara, R Sakamoto, Y Ohta, “Multi-Resolution Bullet-Time Effect”, ACMSIGGRAPH-Asia, 2014

[7] T Kanade et al., “Eye Vision”, http://www.ri.cmu.edu/events/sb35/tksuperbowl.html, access:2017

[8] Replay Technologies, “freeD Technology”, http://replay-technologies.com/, access:2017

[9] Timeslice Films, http://timeslicefilms.com/, access:2017

[10] S Yonemoto, “A Video Annotation Tool Using Vision-based AR Technology”, International Conference on Cyberworlds, pp.226-230, 2012

[11] J Frahm, P Fite-Georgel, D Gallup, T Johnson, R Raguram, C Wu, Y Jen, E Dunn, B Clipp, S Lazebnik, M Pollefeys, “Building Rome on a Cloudless Day”,Computer Vision – ECCV 2010, vol6314, pp.368-381, 2010

[12] Y Furukawa, J Ponce, “Accurate, Dense, and Robust Multi-View Stereopsis”, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.32, Issue8, pp.1362-1376,2010

[13] 後藤祐斗，藤吉弘亘，“Inside-out カメラにおける眼球の輻輳量からの注視点距離推定”，研究報告コンピュータビジョンとイメージメディア (CVIM) 177(15), pp.1-8, 2011

[14] C Wu, "Towards Linear-time Incremental Structure From Motion", 3DV, 2013

[15] C Wu, S Agarwal, B Curless, S Seitz, "Multicore Bundle Adjustment", CVPR, 2011

[16] A Tsuji, S Matsuda, K Suzuki, “Interpersonal distance and face-to-face behavior during the therapeutic activities for children with ASD”, In Computers Helping People with Special Needs. Springer International Publishing, 2016