一般社団法人 人工知能学会
Japanese Society for
Artificial Intelligence

人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-055-7
(11/22)

視聴覚統合による動的環境下における3次元再構成の提案

Audio-visual integration for 3D reconstruction under dynamic environments

紺野 隆志[1]* 西田 健次[1] 糸山 克寿[1] 中臺 一博[1,2]
Takashi Konno[1] Kenji Nshida[1] Katsutoshi Itoyama[1] Kazuhiro Nakadai[1,2]

[1] 東京工業大学
[1] Tokyo Institute of Technology
[2] (株) ホンダ・リサーチ・インスティチュート・ジャパン
[2] Honda Research Institute Japan Co., Ltd.

Abstract: Structure from motion (SfM) usually assumes a stationary environment where no moving object exists. In practice, the environment is dynamic where many moving objects exist. It makes the performance of SfM poor. In this paper, to solve this problem, we focus on the fact that dynamic objects often emit sound, and propose an audiovisual 3D environment understanding method that integrates acoustic signal processing by microphone array processing into SfM. The performance of the proposed method is evaluated on a public dataset which simulates dynamic environments.

1 はじめに

3次元再構成は、ARやVR、ロボティクスなど様々な応用先があり、コンピュータビジョンにおける最も重要な分野の1つとして、多くのアルゴリズムが提案されている。例えば過去20年間で、Structure from Motion (SfM) [1] や Muti-View Stereo (MVS) [2] 、Simultaneous Localization and Mapping (SLAM) [3, 4] などが提案されている。

SfMは、物体やシーンに対して様々な視点で撮影した2次元の画像群から、カメラの位置と姿勢および、物体の3次元構造を復元する手法である [5]。物体が静的であるという仮定のもとで、カメラと物体の幾何学的な計算により3次元構造が復元される。画像内に動的物体が存在している場合は、多くの場合、画像間の特徴点マッチングにより動的物体は除外される。そのため、復元された3次元構造の点群が存在しない領域には、何も物体は存在しないのか、それとも動的物体が存在していたが除外されたのかを判別することはできない。また、特徴点マッチングにおける除外処理に失敗した場合、カメラの姿勢推定の誤りが大きくなり、いびつな形状の物体が復元されてしまう場合がある。この場合、動的物体だけではなく静的な物体の復元性能にも影響を与えてしまう。

*連絡先：東京工業大学 工学院 システム制御系
〒 152-8552 東京都目黒区大岡山 2-12-1
E-mail: itoyama@ra.sc.e.titech.ac.jp

現実世界では、動的物体は、その動作や振動などにより音を発している場合が多い。例えば、走行している車はエンジンの振動音やロードノイズ、風切り音などを発し、歩行している人は足音を発している。これは、従来のSfMでは再構成から除外される動的物体は、音情報を利用することにより再構成ができる可能性があることを示唆している。そこで本稿では、動的物体は常に音を発していると仮定をし、音源定位とSfMを統合した動的環境下における3次元再構成を提案する。音と画像の空間的な対応関係を利用することにより、画像内の静的物体と各動的物体を分け、それぞれの物体ごとにSfMを行い3次元再構成をする。最後に、静的物体と動的物体を統合し、全体構造を復元する。動的物体とその物体が発する音の対応関係がとれていることにより、定位した音の視覚的な3次元構造も確認可能となる。そのため、画像だけを用いて3次元再構成を行うよりも環境理解が深まることが期待される。

2 関連研究

動的環境下における3次元再構成 動的物体の3次元復元は、コンピュータビジョンの研究でも困難な問題とされている。そのため、動的環境下における3次元再構成の多くの従来研究は、動的領域を外れ値として除外することを目指している。外れ値除去アルゴリズムとして、RANSAC [6] が最も用いられている。近年

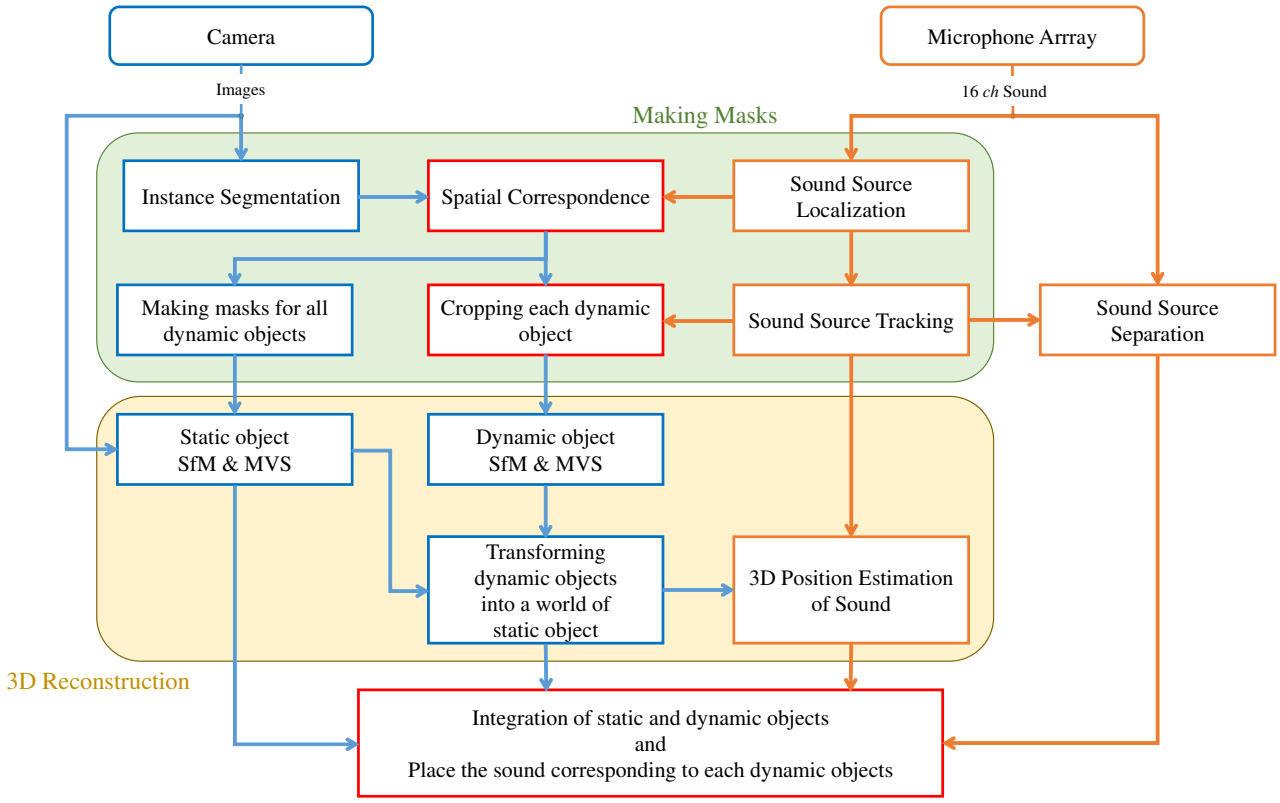


図 1: Flowchart of the proposed system. Blue frame modules indicate processing of images. Orange frame modules indicate the processing of sound. Red frame modules indicate the processing of both image and sound. Blue arrows indicate the flow of information about images. Orange arrows indicate the flow of information about sound. The modules on the green background perform making masks. Those on the yellow background perform 3D reconstruction.

は、深層学習による性能向上を背景に物体検出などを利用した新しい動的物体除去のアルゴリズムが提案されている [7, 8]。RGB 画像だけでなく、Depth 画像も用いた動的物体の 3 次元構造復元アルゴリズムとして、Co-Fusion [9] や MaskFusion [10]、MID-Fusion [11]、DetectFusion [12] などが提案されている。いずれも、本稿と同じく静的物体と動的物体を分離し、それぞれの物体ごとに 3 次元復元を行っている。Depth 画像を利用しているため、RGB 画像のみを利用する SfM よりも復元が容易となり復元性能もよいが、depth の測定可能距離には限界があり、屋外などの直射日光環境では使用が困難である。本稿は、RGB 画像と音を用いることにより、屋外での使用も目指し、さらに音の情報を利用しより高次元な環境理解を目指す。

音情報を利用した 3 次元再構成 マイクロホンアレイ処理を用いた音源定位と、カメラや LiDAR による SLAM を用いて、音源の 3 次元位置を推定し、SLAM により復元したポイントクラウド上に音源を表示する研究が行われている [13, 14, 15, 16]。しかし、いずれの研究も定位結果の三次元空間へのマッピングが目的であり、音響情報と画像情報の相補的な統合をし性能を向上させるという取り組みはされていない。このため、本稿で目指すような、視聴覚情報を統合して、SfM の問題点を解決するという課題に対して、これらの手法を適用することは難しい。

3 提案手法

図 1 に、提案手法のフレームワークを示す。まず、音と画像の空間的な関係を利用し、各画像ごとに各動的物体のバイナリマスクを作成する。音源追跡により、画像間の各動的物体をトラッキングし、全画像の動的物体それぞれに対応するバイナリマスクを得る。次に、このバイナリマスクを用いて、静的物体と各動的物体ごとに SfM と MVS を適用し、それぞれの物体ごとに 3 次元構造を復元する。最後に、静的物体と動的物体を統合し、全体シーンを復元する。別のフローとして、音源定位により得られた音源の空間情報を用いて音源分離を行うことにより、各動的物体に対応する音および

その視覚的な 3 次元構造を得る。各モジュールの詳細については、以降の節で述べる。

次に、カメラとマイクロホンアレイの配置について述べる。カメラとマイクロホンアレイの相対的な位置と姿勢の関係を常に一定に保つため、カメラの上部にマイクロホンアレイを取り付ける。その際、カメラの光軸方向とマイクロホンアレイの 0 度方向が同じ方向を向くようにする。そのため、カメラの動きに合わせてマイクロホンアレイの位置と姿勢も変動する。

3.1 音と画像の空間的な対応関係によるバイナリマスク生成

3.1.1 インスタンスセグメンテーション

全画像 N に対して、インスタンスセグメンテーションを適用し、画像 $\{I_i\}_{i=1}^{N} \in \mathbb{R}^{w\times h\times 3}$ 内に映る物体 $o \in \{1, ..., K\}$ の Boundin Box(BBox) $b_{i,o} \in \mathbb{R}^4$ およびそのバイナリバイナリマスク $M_{i,o} \in \mathbb{R}^{w\times h}$ を得る。w, h は画像の幅と高さであり、K は画像 i において検出される物体数である。インスタンスセグメンテーションのアルゴリズムとして、オフラインで最も性能のよい Mask-RCNN [17] を利用する。検出される物体には、静的な物体も含まれる。

3.1.2 音源定位

全音源数を L とする。マイクロホンアレイ処理を用いた音源定位により、画像 i におけるマイクロホンアレイに対する音源 $s \in \{1, ..., L\}$ の方位角 $\theta_{i,s}$ と仰角 $\phi_{i,s}$ を得る。音源定位のアルゴリズムとして、MUSIC (MUltiple Signal Classification) 法 [18] を利用し、実装にはロボット聴覚 OSS である HARK (Honda Research Institute Japan Audition for Robots with Kyoto University) [19] を用いる。得られた音源の方向とカメラの内部パラメータ $A \in \mathbb{R}^{3\times 3}$ を利用して、音源の 3 次元位置 $P_s \sim [\tan\theta_{\mathrm{i,s}}\cos\phi_{\mathrm{i,s}}, \tan\theta_{\mathrm{i,s}}\sin\phi_{\mathrm{i,s}}, 1]^{\mathrm{T}}$ を画像に投影することにより、音源 s の画像 i 内の位置 $P_{i,s}(\sim AP_s) \in \mathbb{R}^2$ を得る。あらかじめ任意に定めたオフセット off を用いて、以下の式により音源の BBox $b_{i,s} \in \mathbb{R}^4$ を得る。

$$b\_\mathrm{xmin_{i,s}} = \mathrm{P_{i,s}\_x} - \mathrm{off}, \quad b\_\mathrm{ymin_{i,s}} = \mathrm{P_{i,s}\_y} - \mathrm{off} \,(1)$$

$$b\_\mathrm{xmax_{i,s}} = \mathrm{P_{i,s}\_x} + \mathrm{off}, \quad b\_\mathrm{ymax_{i,s}} = \mathrm{P_{i,s}\_y} + \mathrm{off} \,(2)$$

3.1.3 音と画像の空間的な対応関係

画像 i において、インスタンスセグメンテーションにより推定された全 BBox $b_{i,o}$ と、音源定位により推定された全 BBox $b_{i,s}$ から全ペアを抽出し、各ペアの Intersection-over-Union ($\mathrm{IoU_{i,o,s}}$) を計算する。IoU が任意のしきい値 th_{iou} を超えた場合は、そのペアの $b_{i,o}$ は音源つまり動的物体の BBox であるとする。この動的物体のバイナリマスクとして、物体 o に対するバイナリマスク $M_{i,o}$ を用いる。いずれの音源の BBox $b_{i,s}$ とも IoU がしきい値 th_{iou} を超えなかった BBox $b_{i,o}$ は、静的な物体である可能性が高いため、この物体のバイナリマスク $M_{i,o}$ は後の処理では使用しない。しかし、いずれの BBox $b_{i,o}$ とも IoU がしきい値 th_{iou} を超えなかった音源の BBox $b_{i,s}$ は、動的物体の可能性が高いが、インスタンスセグメンテーションによるバイナリマスクは得られない。そこで、この音源の BBox $b_{i,s}$ に含まれる領域を動的物体のマスクとするバイナリマスク $M_{i,s} \in \mathbb{R}^{w\times h}$ を生成し、静的な物体の復元のみに使用する。上記より、画像 i における音源 s に対応する動的物体のバイナリマスク $M_i^s \in \mathbb{R}^{w\times h}$ は以下のように再定義される。

$$M_i^s \leftarrow \begin{cases} M_{i,o} & \text{if} \quad \mathrm{IoU_{i,o,s}} \geq \mathrm{th_{iou}} \\ M_{i,s} & \text{if} \quad \text{otherwise} \end{cases} \tag{3}$$

3.1.4 全動的物体に対するバイナリマスクの生成

静的物体の復元の際に使用する、全動的物体に対するバイナリマスクの生成について述べる。画像 i における全動的物体のマスクをすべて含むように、以下のように画像 i におけるバイナリマスク $M_i \in \mathbb{R}^{w\times h}$ を生成する。m は、M_i^s と同次元で各値が 1 の行列である。

$$M_i = \left(Lm^{-1}\right) \sum_{s}^{L} M_i^s \tag{4}$$

3.1.5 音源追跡

音源 s を音源追跡することにより、対応する動的物体を画像間でトラッキングし、式 (5) に示す全画像の各動的物体に対応するバイナリマスク群 $M^s \subset \mathbb{R}^{w\times h}$ を得る。音源追跡のアルゴリズムとして、HARK の SourceTracker[1] を利用する。

$$M^s = \{M_i^s \mid i = 1 \ldots N\} \tag{5}$$

3.1.6 各動的物体のみが映った画像の生成

各動的物体の復元の際に使用する、各動的物体のみが映った画像の生成について述べる。全画像に対して各動的物体に対応するバイナリマスクを掛けあわせる

[1] https://www.hark.jp/document/2.0.0/hark-document-ja/subsec-SourceTracker.html

ことにより、以下のように音源 s に対応する動的物体のみが映った画像群 $\boldsymbol{D}^s \subset \mathbb{R}^{w \times h \times 3}$ を生成する。

$$\boldsymbol{D}^s = \{\boldsymbol{D}_i^s \mid \boldsymbol{D}_i^s = M_i \times I_i, \quad i = 1 \ldots N\} \tag{6}$$

3.2 静的物体と動的物体の３次元構造復元

3.2.1 静的物体の復元

画像 i と対応する全動的物体に対するバイナリマスク M_i をペア (I_i, M_i) として、全ペアを SfM と MVS へと入力し、各カメラ姿勢と静的物体の３次元構造を復元する。SfM の処理の際に、バイナリマスクによりマスクされる領域からは特徴点を抽出しないようにし、動的物体を除外する。動的物体を除外することにより、性能向上が期待される。上記の処理の実装のベースとして、OSS の COLMAP [20] を用いる。

3.2.2 動的物体の復元

静的物体と動的物体が両方画像内に映っている場合、多くの場合 SfM の特徴点マッチングにおける幾何学的な外れ値処理により、動的物体は除外されてしまう。そこで、本稿では画像から動的物体のみ抽出して、動的物体のみが映った画像群を新しく生成する。この画像群においては、動的物体が剛体の場合は、擬似的に静的物体とみなすことができるため、SfM によって復元が可能となる。そのため、セクション 3.1.6 によって生成した音源 s に対応する動的物体のみが映った画像群 $\boldsymbol{D}^s$ を SfM に入力することにより、各動的物体のみの３次元構造が復元可能となる。

3.2.3 各動的物体を静的物体のワールドへ変換

SfM では、物体は任意のスケールで復元されるため、動的物体の復元物のワールド (DW) と静的物体の復元物のワールド (SW) は、それぞれワールド座標系が異なる。そのため、各動的物体を静的物体のワールドへ変換する必要がある。

動的物体に対する相対的なカメラ位置と姿勢は、DW と SW でスケールを除き共通である。そのため、カメラ座標系を介することにより動的物体を、DW のワールド座標系に対する３次元位置 ${}^{\text{world}}P_{i,DW}^s$ から SW のワールド座標系に対する３次元位置 ${}^{\text{world}}P_{i,SW}^s$ へと変換する。

まず、式 (7) により、動的物体を DW におけるワールド座標系からカメラ座標系へ変換する。DW におけるワールド座標系からカメラ座標系への回転行列を $R_{DW} \in \mathbb{R}^{3\times3}$、並進行列 $T_{DW} \in \mathbb{R}^3$ と表す。

$${}^{\text{cam}}P_{i,DW}^s = R_{DW} \times {}^{\text{world}}P_{i,DW}^s + T_{DW} \tag{7}$$

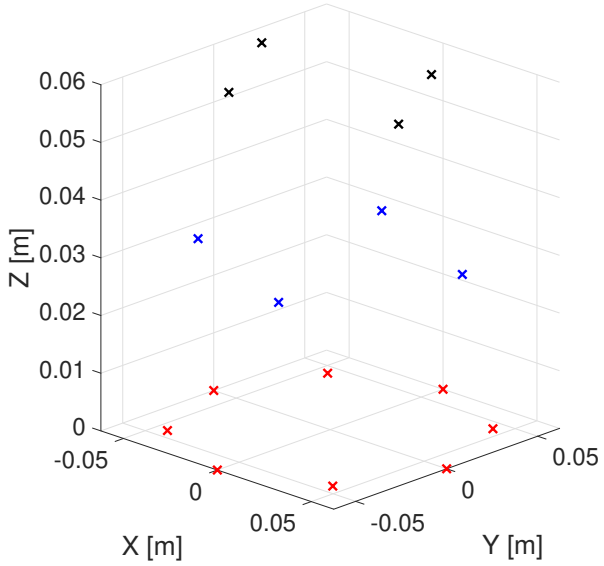


図 2: Microphone configuration of Microphone Array.

次に、式 (8) により、動的物体を DW におけるカメラ座標系 ${}^{\text{cam}}P_{i,DW}^s$ から、SW におけるカメラ座標系 ${}^{\text{cam}}P_{i,SW}^s$ へ変換する。DW から SW へのスケール変換を $S_{DW2SW} \in \mathbb{R}$ と表す。

$${}^{\text{cam}}P_{i,SW}^s = S_{DW2SW} \times {}^{\text{cam}}P_{i,DW}^s \tag{8}$$

最後に、式 (9) により、動的物体を SW におけるカメラ座標系 ${}^{\text{cam}}P_{i,SW}^s$ からワールド座標系 ${}^{\text{world}}P_{i,SW}^s$ へ変換する。SW におけるワールド座標系からカメラ座標系への回転行列を $R_{SW} \in \mathbb{R}^{3\times3}$、並進行列 $T_{SW} \in \mathbb{R}^3$ と表す。

$${}^{\text{world}}P_{i,SW}^s = R_{SW}^{-1} \times ({}^{\text{cam}}P_{i,SW}^s - T_{SW}) \tag{9}$$

3.3 全体シーンの復元および音源分離

式 (9) により、SW における画像 i に対する音源 s に対応する動的物体の３次元位置 ${}^{\text{world}}P_{i,SW}^s$ が得られる。画像 i に対応する時刻 t において、SW の ${}^{\text{world}}P_{i,SW}^s$ に各動的物体を配置することにより、時間的に変動する３次元構造が復元される。${}^{\text{world}}P_{i,SW}^s$ に、音源分離により分離した音源 s の音を配置することにより、各動的物体に対応する音およびその視覚的な３次元構造を得る。音源分離のアルゴリズムとして、HARK に実装されている GHDSS[2] (Geometric High-order Dicorrelation-based Source Separation) を用いる。

[2] https://www.hark.jp/document/2.0.0/hark-document-ja/subsec-GHDSS.html

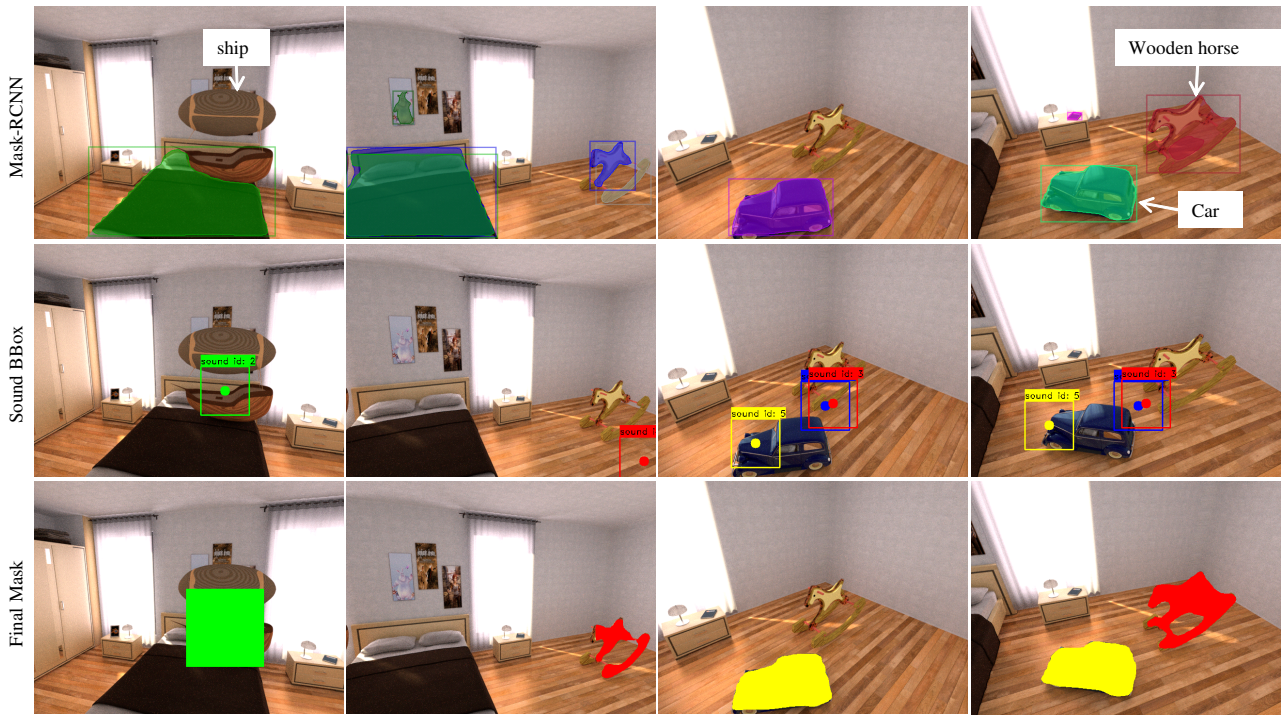
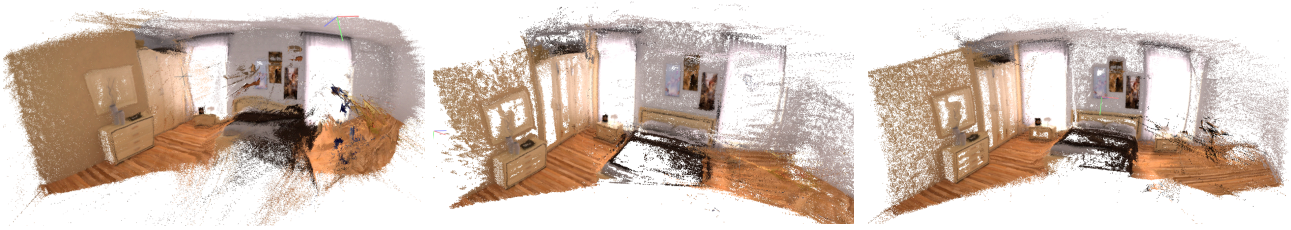


図 3: Qualitative results for making binary masks of dynamic objects. Color-coded for each detected object and sound source.



(a) Without Masks (b) With Predicted Masks (c) With Ground Truth Masks

図 4: Qualitative results for 3D reconstruction of static objects

4 評価実験

動的物体の 3 次元構造復元の評価を目的として、Martin らによって作成された Co-Fusion データセット [9] を用いて、提案手法の定性評価を実施し、手法の有効性および現状の限界を示す。評価実験として、動的物体のバイナリマスクの性能評価 (4.2)、静的物体の復元性能評価 (4.3)、動的物体の復元性能評価 (4.4)、全体シーンの復元性能評価 (4.5) を実施した。

4.1 実験設定

Co-Fusion データセット：Co-Fusion データセットには、複数の物体 (静的物体と動的物体いずれも) が存在する環境でカメラを動かして撮影した画像 (RGB 画像と Depth 画像) や、各時刻におけるカメラや動的物体の 3 次元位置の真値などが含まれている。シミュレーション環境と実環境で取得した、合計 4 つの環境でのデータが含まれる。本稿では、シミュレーション環境における 850 枚の RGB 画像を使用した。シミュレーションで再現した部屋の中に、3 つの動的物体 (Ship, Wooden Horse, Car) がそれぞれ独立して動いており、常に画像内に動的物体が写っているとは限らない。

音のシミュレーション：Co-Fusion データセットには、音が含まれていないため、シミュレーションで音を再現した。動的物体は常に音を発していると仮定し、各時刻における各動的物体の 3 次元位置の真値に音源を置いた。音は各動的物体の見た目に合わせて、16.1[kHz] で録音されたモノラル音を用いた。音の録音には、16 ch のマイクロホンアレイを用い、0 度方向がカメラの光軸方向と合うようにカメラに固定した。16 個のマイ

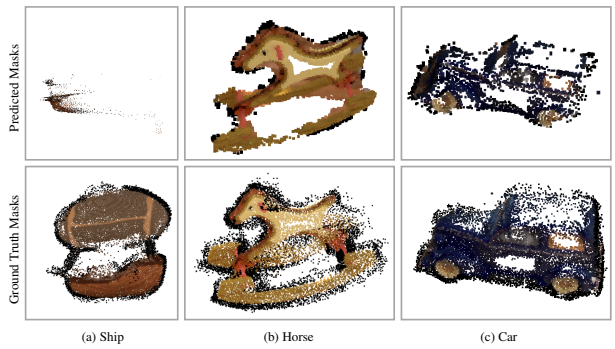


図 5: Qualitative results for 3D reconstruction of dynamic objects.

クロホンは図 2 のように、最下段に 8 個、高さ 3 cm の中段に 4 個、高さ 6 cm に 4 個配置した。音源定位には、このマイクロホンアレイに対して幾何的に計算した伝達関数を用いた。実際は音源とマイクロホンアレイどちらも動いているが、マイクロホンアレイは固定し音源を相対的に動かした。各フレームにおいて各マイクロホンと各音源の伝達関数を作成し、そのフレームの音に畳み込み、すべての音源の音を足し合わせることにより 16 ch の混合音を作成した。この混合音を用いて、システムの評価を行った。

インスタンスセグメンテーション：Mask-RCNN は、Detectron2 [21] で実装されているコードを利用し、ResNet-101 [22] と FPN をバックボーンとし MS COCO データセットの train2017 で学習済みのモデルを使用した。

4.2 動的物体のバイナリマスクの性能評価

図 3 に、Mask-RCNN(図の 1 段目) と Sound BBox(図の 2 段目) により動的物体のバイナリマスク (図の 3 段目) を生成した結果を示す。Ship は、学習済みモデルに含まれていないため Mask-RCNN では検出されない。そのため、4.2 で示した方法で音を用いてバイナリマスクを生成しているが、Ship 全体を覆うマスクは生成できていない。Horse と Car については、ある程度精度よくバイナリマスクを生成できている。しかし、音源間の距離が近づく場合に、音源定位の分解能の限界によりうまく二つの音源を定位することができず、音源追跡に失敗している場合があった。

4.3 静的物体の復元性能評価

図 4 に、静的物体の復元結果を示す。(a) は動的物体のバイナリマスクなし、(b) は提案手法により推定したバイナリマスクあり、(c) は Ground Truth のバイナリマスクありで、それぞれ SfM と MVS により復元した結果である。(a) は、動的物体が存在している領域に歪みが生じて復元されている。動的物体のマスクを使用しないため、画像間のマッチングで動的物体の特徴点除去に失敗し、カメラ姿勢推定誤差が大きくなってしまうためと考えられる。提案手法では 4.2 で示した通り完全なマスクを推定することができていないが、(b) の結果から (a) で見られる歪みをある程度抑えられていることが確認できる。さらに、動的物体を完全にマスクした (c) の復元結果に近い結果が得られている。完全ではないものの動的物体の特徴点をある程度除去することができているため、画像間マッチングの除去処理がうまく働いたと考えられる。

4.4 動的物体の復元性能評価

図 5 に、各動的物体の復元結果を示す。1 段目は提案手法、2 段目は Ground Truth のバイナリマスクを用いて復元した結果である。Ground Truth のマスクを用いた場合でも、画像から動的物体のみを抽出することにより画素数が小さく、動的物体の特徴点数が少ないため若干歪みが生じている。提案手法では、Ship は 4.2 の通りマスクの性能がよくなく、Ship 全体を覆うマスクではないため、全体を復元することはできていない。そのため Ship のマスクは、静的物体の復元に影響を与えないように生成することが主な目的となる。Horse と Car については、ある程度よく復元ができているが、マスクが生成できていないフレームもあり、Ground Truth よりも復元に使用する画像数が少なくなり、復元される点群数が少ない。

4.5 全体シーンの復元性能評価

図 6 に、動的物体に対する Ground Truth のバイナリマスクを用いて、全体シーンを復元した結果を示す。1 段目は元の画像、2 段目は 1 段目に対応する時間に復元されたポイントクラウドを元画像と同じサイズの空の画像に投影した結果、3 段目は 1 段目に対応する時間に復元されたポイントクラドを上から見た図である。3.2.2 と 3.2.3 の性能評価を行うため、動的物体に対するバイナリマスクは Ground Truth を用いた。ある程度よく、動的物体を DW から SW へ変換することができているが、4.4 で述べたように、動的物体の復元性能 (カメラ姿勢推定) があまりよくないため、各時刻で動的物体が振動していたり位置がずれてしまっている。また、S_{DW2SW} は任意に決めているため、自動で推定する方法の開発は Future work である。

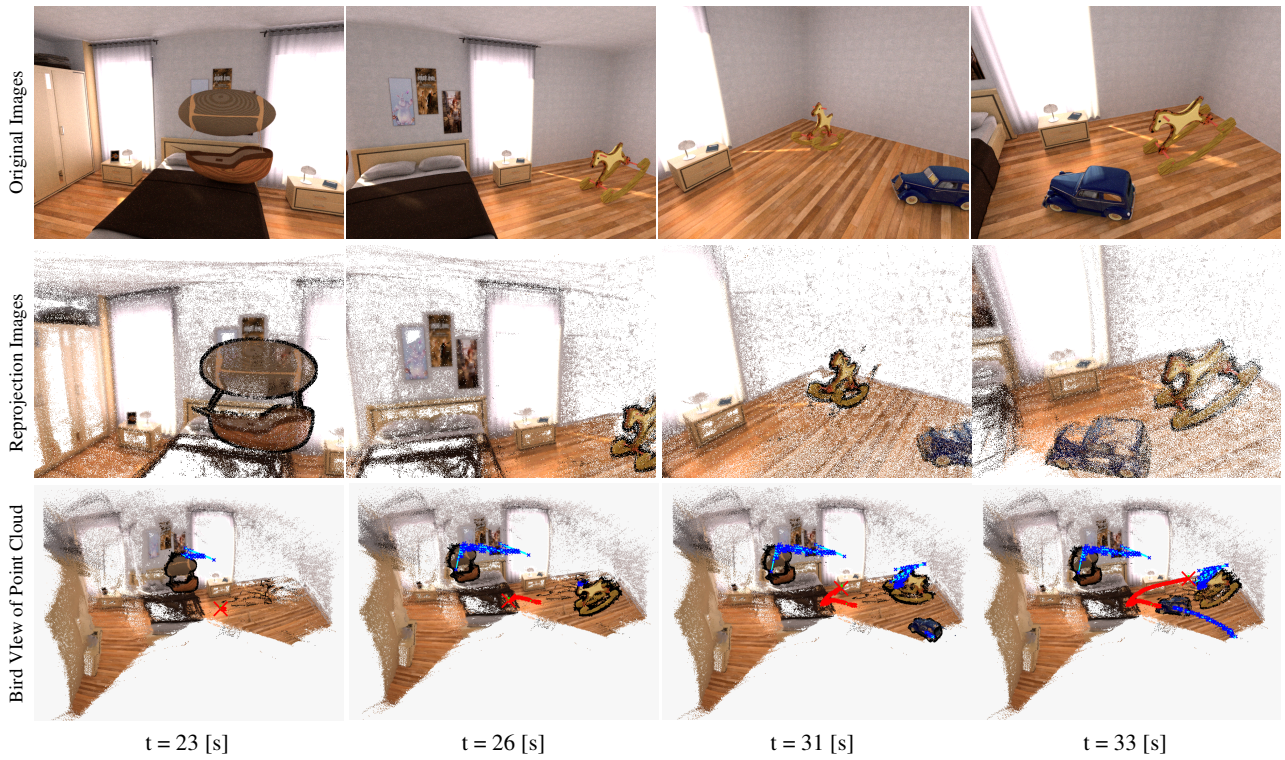


図 6: Qualitative results for 3D reconstruction of all scenes. The red line in the third row figure represents the camera trajectory, and the blue line represents the dynamic object trajectory.

5 おわりに

本稿では、SfM ではうまく再構成ができない動的環境下において、音響信号を手がかりに 3 次元再構成を行う手法について述べた。Co-Fusion データセットを用いて提案手法の定性評価を実施し、音響信号を用いることによって、SfM の性能を向上できる可能性があることを示した。複数の音源が近くに存在する場合に音源定位がうまくいかなくなる場合や、動的物体のスケールに関する問題など、本手法の限界も示した。今後は、設定が異なるシミュレーション環境や実環境、屋外などでの評価実験を行う。また、画像内に動的物体が写っていない場合の手法や、画像と音を統合した音源分離の手法などに取り組み、画像と音を両方利用することによる 3 次元環境理解の有効性を示していく。

謝辞

音のシミュレーションに関して助言を頂いた、山田泰基氏と鍾知氏には深く感謝したします。

参考文献

[1] S. Agarwal, N. Snavely, I. Simon, S. M. Seitz, and R. Szeliski. Building rome in a day. In *IEEE International Conference on Computer Vision (ICCV)*, 2009.

[2] Y. Furukawa and J. Ponce. Accurate, dense, and robust multi-view stereopsis. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition(CVPR)*, 2007.

[3] J. Engel, T. Schöps, and D. Cremer. Lsd-slam: Large-scale direct monocular slam. In *IEEE European Conference on Computer Vision (ECCV)*, 2014.

[4] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos. Orb-slam: A versatile and accurate monocular slam system. *IEEE Transactions on Robotics*, 31(5), 2015.

[5] R. Hartley and A. Zisserman. Multiple view geometry in computer vision., 2004. Cambridge University Press.

[6] M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–3951, 1981.

[7] Bescos Berta, Fácil JM., Civera Javier, and Neira José. DynaSLAM: Tracking, mapping and inpainting in dynamic environments. *IEEE Robotics and Automation Letters*, 2018.

[8] F. Zhong, S. Wang, Z. Zhang, C. Chen, and Y. Wang. Detect-slam: Making object detection and slam mutually beneficial. In *2018 IEEE Winter Conference*

on Applications of Computer Vision (WACV), pages 1001–1010, 2018.

[9] Martin Rünz and Lourdes Agapito. Co-fusion: Real-time segmentation, tracking and fusion of multiple objects. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 4471–4478, 2017.

[10] M. Runz, M. Buffier, and L. Agapito. Maskfusion: Real-time recognition, tracking and reconstruction of multiple moving objects. In *IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 10–20, 2018.

[11] Binbin Xu, Wenbin Li, Dimos Tzoumanikas, Michael Bloesch, Andrew Davison, and Stefan Leutenegger. Mid-fusion: Octree-based object-level multi-instance dynamic slam. In *IEEE International Conference on Robotics and Automation*, International Conference On Robotics and Automation. IEEE, 5 2019.

[12] Ryo Hachiuma, Christian Pirchheim, Dieter Schmalstieg, and Hideo Saito. Detectfusion: Detecting and segmenting both known and unknown dynamic objects in real-time slam. In *British Machine Vision Conference (BMVC)*, 2019.

[13] Y. Sasaki, R. Tanabe, and H. Takemura. Probabilistic 3d sound source mapping using moving microphone array. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1293–1298, 2016.

[14] D. Su, T. Vidal-Calleja, and J. V. Miro. Towards real-time 3d sound sources mapping with linear microphone arrays. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 1662–1668, 2017.

[15] D. Su, K. Nakamura, K. Nakadai, and J. V. Miro. Robust sound source mapping using three-layered selective audio rays for mobile robots. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2771–2777, 2016.

[16] D. Su, T. Vidal-Calleja, and J. V. Miro. Split conditional independent mapping for sound source localisation with inverse-depth parametrisation. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2000–2006, 2016.

[17] He Kaiming, Gkioxari Georgia, Dollar Piotr, and Girshick Ross. Mask r-cnn. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[18] R. Schmidt. Multiple emitter location and signal parameter estimation. *IEEE Transactions on Antennas and Propagation*, 34(3):276–280, 1986.

[19] K. Nakadai, T. Takahashi, H. G. Okuno, H. Nakajima, Y. Hasegawa, and H. Tsujino. Design and implementation of robot audition system 'hark '–open source software for listening to three simultaneous speakers. *Advanced Robotics*, 24:739–761, 2010.

[20] Schönberger, Johannes Lutz, Frahm, and Jan-Michael. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[21] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. `https://github.com/facebookresearch/detectron2`, 2019.

[22] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.