Audio-Visual SfM towards 4D reconstruction under dynamic scenes

Takashi Konno[1] Kenji Nishida[1] Katsutoshi Itoyama[1] Kazuhiro Nakadai[1,2]
[1] Tokyo Institute of Technology, Japan* [2] Honda Research Institute Japan Co., Ltd., Japan

1. Introduction

3D reconstruction is currently one of the most important and popular issues in computer vision, consequently many algorithms have been studied in the last twenty years. SfM is a high performant method to estimate the 3D structure of an object from 2D images taken by a camera from various viewpoints, with estimation of the positions and poses of the camera.Since it assumes everything is stationary in a scene, dynamic objects , which may be moving, vibrating, or rotating, are eliminated at the feature point matching stage of SfM. If such objects are eliminated from the reconstructed 3D structure, one cannot know whether they were present but invisible due to feature point matching error, or absent. If such objects are not eliminated and reconstructed with deformations, their inclusion may adversely affect reconstruction of the whole scene. SfM also has a problem that it cannot handle time information and can only perform 3D reconstruction. For example, information on the movement of dynamic objects cannot be reconstructed. In recent years, with the improvement of object recognition accuracy by deep learning, there have been proposed methods [4, 5, 8, 1] for 3D reconstruction of dynamic scenes. However, there are problems such as restrictions on the application range and complicated processing / calculation amount.

In the real world, dynamic objects usually emit sounds, for example, the rustling of tree leaves on a windy day, and the footsteps of a walking/running person. This suggests that a dynamic object usually eliminated from reconstruction with conventional SfM may be structured using sound information. This paper proposes a method to reconstruct a dynamic object by integrating visual SfM with sound source localization, which we call "Audio-Visual SfM," on the assumption that a dynamic object tends to emit sound. First, static objects and each dynamic objects in the images are separated by utilizing a spatial correspondence between a sound and an image. For each separated object, the 3D structure is reconstructed by performing SfM. Finally, the static and dynamic objects are integrated using the time series information of the sound, and the 4D reconstruction including the time information is performed. Since the correspondence between the dynamic object and the sound emitted by the object is taken, it is possible to visualize the 3D structure of a sound source object.

*{konno, nishida, itoyama, nakadai}@ra.sc.e.titech.ac.jp

2. Overview of Proposed Method

Figure 1 shows the framework of the proposed method. First, a binary mask of each dynamic objects is made for each image using the spatial relationship between sound and images. With sound source tracking, each dynamic object between images is tracked, and a binary mask corresponding to each dynamic object in all images is obtained. Next, using this binary mask, SfM and MVS are applied to the static object and each dynamic object, and the 3D structure is reconstructed for each objects. Finally, the static and dynamic objects are integrated to reconstruct the entire scene. As another flow, by performing sound source separation using spatial information of a sound source obtained by sound source localization, a sound corresponding to each dynamic object and its visual 3D structure are obtained.

The system is constructed assuming the following arrangement of the camera and microphone array. The microphone array is mounted on the top of the camera to keep the relationship between the relative pose of the camera and the microphone array. At this time, the direction of the optical axis of the camera and the direction of 0 degrees of the microphone array are oriented in the same direction. Therefore, the pose of the microphone array also change in accordance with the movement of the camera.

2.1. Separate Static and Dynamic Objects

Separate static and dynamic objects using binary mask for dynamic objects. Utilizing the spatial relationship between sound and images, a binary mask for each dynamic object is made for each image. First, a region that is a candidate for a dynamic object is detected using instance segmentation and sound source localization. The outputs of the two methods are integrated to form a final masks. When the intersection of the two methods of the BBox is large, the mask of the BBox is used as the mask of the dynamic object. The mask estimated for each frame is tracked for all frames using sound source tracking.

2.1.1 Instance segmentation

Apply instance segmentation to the entire image N, and get the Boundin Box (BBox) $b_{i,o} \in \mathbb{R}^4$ of the objects $o \in \{1, ..., K\}$ in the image $\{I_i\}_{i=1}^{N} \in \mathbb{R}^{w \times h \times 3}$ and its binary binary mask $M_{i,o} \in \mathbb{R}^{w \times h}$. w, h are the width and height of the image, and K is the number of objects detected

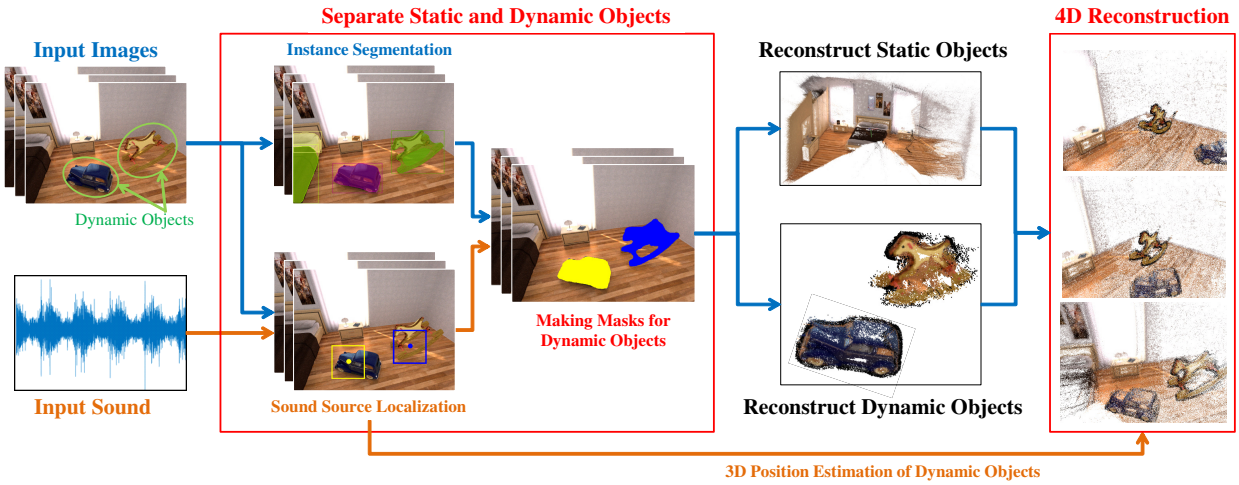


Figure 1. Flowchart of proposed Audio-Visual SfM

in the image i. As the algorithm for instance segmentation, we use the Mask-RCNN [2]. The detected object includes a static object.

2.1.2 Sound source localization

Let the total number of sound sources be L. By sound source localization using microphone array processing, the azimuth $\theta_{i,s}$ and elevation $\phi_{i,s}$ of the sound source $s \in \{1, ..., L\}$ with respect to the microphone array in the image i are obtained. The MUSIC (MUltiple Signal Classification) method [6] is used as a sound source localization algorithm. HARK (Honda Research Institute Japan Audition for Robots with Kyoto University) [3], which is a robot auditory OSS, is used for implementation. Using the obtained sound source direction and camera internal parameters $A \in \mathbb{R}^{3\times3}$, By projecting 3D position of sound source $P_s \sim [\tan\theta_{\mathrm{i,s}}\cos\phi_{\mathrm{i,s}}, \tan\theta_{\mathrm{i,s}}\sin\phi_{\mathrm{i,s}}, 1]^{\mathrm{T}}$ onto the image, the 2D position $P_{i,s}(\sim AP_s) \in \mathbb{R}^2$ of the sound source s in the image i is obtained. Using an arbitrary predetermined offset off, the sound source BBox $b_{i,s} \in \mathbb{R}^4$ is obtained by the equation (1).

$$b_{i,s} = \begin{bmatrix} x_{\min} \\ y_{\min} \\ x_{\max} \\ y_{\max} \end{bmatrix} = \begin{bmatrix} P_{i,s,x} - \mathrm{off} \\ P_{i,s,y} - \mathrm{off} \\ P_{i,s,x} + \mathrm{off} \\ P_{i,s,y} + \mathrm{off} \end{bmatrix} \tag{1}$$

2.1.3 Making Masks for Dynamic Objects

In image i, all pairs are extracted from all BBox $b_{i,o}$ estimated by instance segmentation and all BBox $b_{i,s}$ estimated by sound source localization. Calculate Intersection-over-Union ($\mathrm{IoU_{i,o,s}}$) of each pair. When the IoU exceeds an arbitrary threshold value th_{iou}, it is assumed that $b_{i,o}$ of the pair is a sound source, that is, a BBox of a dynamic object. As a binary mask of the dynamic object, a binary mask $M_{i,o}$ for the object o is used. BBox $b_{i,o}$, whose IoU dose not exceed the threshold th_{iou} for any BBox of sound source $b_{i,s}$, is likely to be a static object. Therefore, the binary mask $M_{i,o}$ of this object is not used in the subsequent processing. However, the BBox $b_{i,s}$ of the sound source whose IoU did not exceed the threshold value th_{iou} for any BBox $b_{i,o}$ is likely to be a dynamic object. However, the binary mask by instance segmentation cannot be obtained. Therefore, the binary mask $M_{i,s} \in \mathbb{R}^{w\times h}$ using the area included in BBox $b_{i,s}$ of this sound source as the mask of the dynamic object. And use it only for reconstructing static objects. From the above, the binary mask $M_i^s \in \mathbb{R}^{w\times h}$ of the dynamic object corresponding to the sound source s in the image i is redefined by the equation (2).

$$M_i^s \leftarrow \begin{cases} M_{i,o} & \text{if } \mathrm{IoU_{i,o,s}} \geq \mathrm{th_{iou}} \\ M_{i,s} & \text{otherwise} \end{cases} \tag{2}$$

2.2. 3D reconstruction of static and dynamic objects

Using the binary masks M_i^s for dynamic objects, the 3D structure of static and dynamic objects are reconstructed separately.

2.2.1 Reconstruct static objects

The binary mask $M_i \in \mathbb{R}^{w\times h}$ including all dynamic objects in the image i is made by the equation (3).

$$M_i = \bigcup_s^L M_i^s \tag{3}$$

As a pair (I_i, M_i) of I_i and M_i, all pairs are input to SfM and MVS, reconstruct the 3D structure of only static objects. During SfM processing, feature points are not extracted from regions masked by the binary mask, and dynamic objects are excluded. OSS COLMAP [7] is used as a base for implementing the above processing.

2.2.2 Reconstruct dynamic objects

When both a static object and a dynamic object appear in an image, a dynamic object is often excluded by geometric outlier processing in SfM feature point matching. Therefore, we extract only the dynamic objects from the image and generate a new image group showing only the dynamic object. By multiplying all the images by the binary mask corresponding to each dynamic object, an image group $\boldsymbol{D}^s \subset \mathbb{R}^{w \times h \times 3}$ including only the dynamic object corresponding to the sound source s is made as follows.

$$\boldsymbol{D}^s = \{\boldsymbol{D}_i^s \mid \boldsymbol{D}_i^s = M_i^s \times I_i,\ \ i = 1 \ldots N\} \tag{4}$$

In this group of images, when the dynamic object is a rigid body, it can be regarded as a pseudo static object, and thus the dynamic object can be reconstructed by SfM.

2.3. 4D Reconstruction

In SfM, since an object is reconstructed at an arbitrary scale, the world of a reconstructed dynamic object (DW) and the world of a reconstructed static object (SW) have different world coordinate systems. Therefore, it is necessary to transform each dynamic object into a world of static objects. We use the fact that the camera pose relative to the dynamic object are the same for DW and SW except the scale. Through the camera coordinate system, the dynamic object is transformed from the 3D pose ${}^{\text{world}}P^s_{i,DW}$ with respect to the world coordinate of DW to the 3D pose ${}^{world}P^s_{i,SW}$ with respect to the world coordinate of SW.

First, the dynamic object is transformed from the world coordinate system in the DW to the camera coordinate system by the equation (5). The rotation matrix and the translation matrix from the world coordinate system to the camera coordinate system in the DW is denoted as $R_{DW} \in \mathbb{R}^{3\times3}$ and $T_{DW} \in \mathbb{R}^3$.

$${}^{\text{cam}}P^s_{i,DW} = R_{DW} \times {}^{\text{world}}P^s_{i,DW} + T_{DW} \tag{5}$$

Next, it is transformed from the camera coordinate system ${}^{\text{cam}}P^s_{i,DW}$ in DW to the camera coordinate system ${}^{cam}P^s_{i,SW}$ in SW by the equation (6). The scale transformation from DW to SW is denoted as $S_{DW2SW} \in \mathbb{R}$.

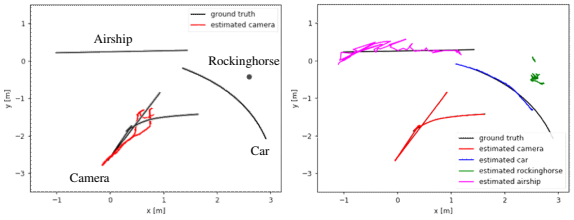
$${}^{\text{cam}}P^s_{i,SW} = S_{DW2SW} \times {}^{\text{cam}}P^s_{i,DW} \tag{6}$$

Finally, it is transformed from the camera coordinate system ${}^{cam}P^s_{i,SW}$ in SW to the world coordinate system ${}^{world}P^s_{i,SW}$ by the equation (7). The rotation matrix and the translation matrix from the world coordinate system to the camera coordinate system in SW is denoted as $R_{SW} \in \mathbb{R}^{3\times3}$ and $T_{SW} \in \mathbb{R}^3$.

$${}^{\text{world}}P^s_{i,SW} = R_{SW}^{-1} \times ({}^{\text{cam}}P^s_{i,SW} - T_{SW}) \tag{7}$$

Table 1. AT-RMSE [cm]

	Original SfM	Ours
Camera	37.49	0.99
Airship	-	40.40
Car	-	21.92
Rockinghorse	-	32.52



(a) Original SfM (b) Ours

Figure 2. Comparison between the ground truth and estimated trajectories for each of the objects.

3. Experiment

We evaluate the proposed method, using the Co-Fusion dataset [4] made by Martin et al. The performance of the reconstruction of the whole scene was evaluated.

Co-Fusion dataset contains images taken by moving the camera in an environment where multiple objects (both static and dynamic objects) exist. The ground truth of the 3D pose of the camera and the dynamic object at each time are included. In this paper, we used 850 RGB images in a simulation environment." In the room made by the simulation, three dynamic objects ("Airship", "Rockinghorse", "Car") move independently, and the dynamic objects are not always shown in the image.

Since the Co-Fusion dataset does not contain sound, the sound was made by simulation. Assuming that the dynamic object always emits sound, the sound source was placed at the ground truth of the 3D pose of each dynamic object.The sound was a monaural sound recorded at 16.1 [kHz]. A 16-channel microphone arraywas used for sound recording. For the sound source localization, a transfer function geometrically calculated for this microphone array was used. In each frame, the transfer function of each microphone and each sound source was made, convolved with the sound of that frame, and the sounds of all sound sources were added to make a 16-channel mixed sound.

3.1. Evaluation of whole scene reconstruction

Table 1 shows the results of the mean square error (RMSE) of Absolute Trajectory (AT)[1] for the orbit of the 3D pose of the estimated camera and the dynamic objects. Fig. 3 shows the result of visualizing the estimated and ground truth trajectories. We compared the results of the original SfM (COLMAP) and the proposed method. In original

[1] https://vision.in.tum.de/data/datasets/rgbd-dataset/tools

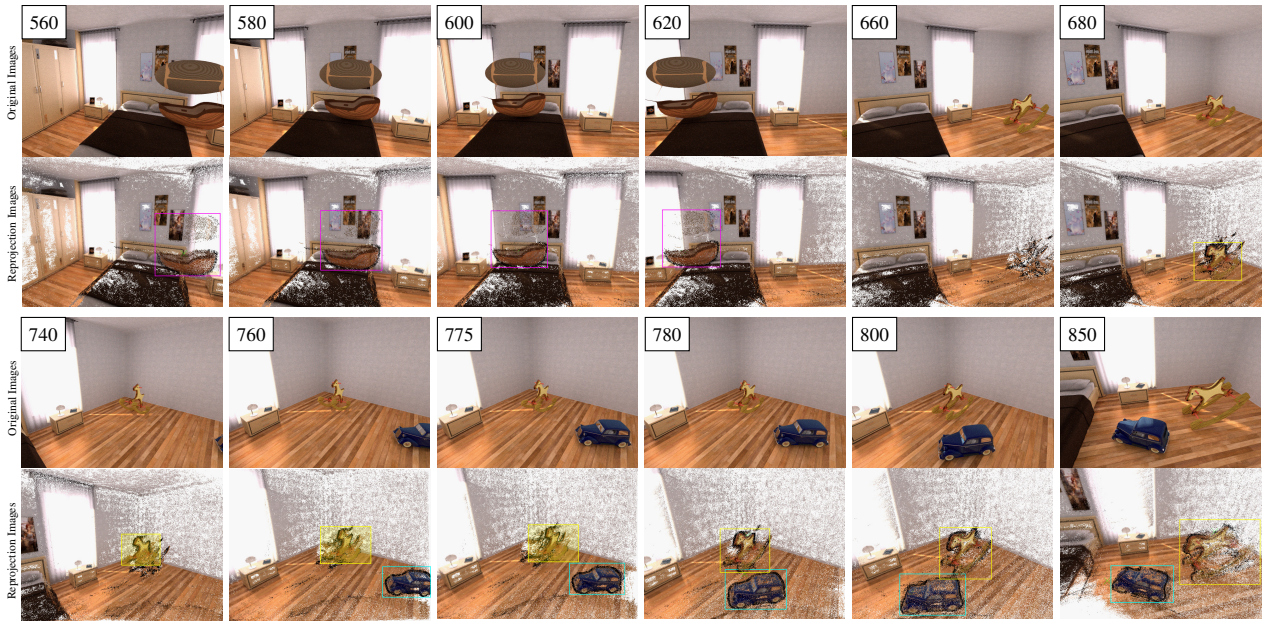


Figure 2. Qualitative results for 3D reconstruction of all scenes. The upper left number indicates the frame. The reconstructed images in the 2nd and 4th rows correspond to their actual images in the 1st and 3rd lines.

SfM, dynamic objects cannot be reconstructed, then only the camera pose is estimated. Comparing the two methods with respect to the camera pose, the original SfM has a larger error due to the influence of the dynamic object than the proposed method. The estimated trajectory of the dynamic objects in the proposed method has some errors, but it can be estimated to some extent accurately.

Figure 2 shows the qualitative result of reconstructing the entire scene. The second and fourth rows show the results of projecting the reconstructed point cloud corresponding to the first and third rows onto a empty image of the same size as the original image. The BBox surrounding the reconstructed dynamic object is also visualized. From the figure, it can be confirmed that the size, the 3D position, the pose of the dynamic object can be almost accurately estimated. However, a slight displacement has occurred. There is also the effect of errors in the estimated camera pose when reconstructing static objects, the reconstruction of a dynamic object is not as accurate as the reconstruction of a static object. In the 660th frame, Rockinghorse could not be reconstructed because mask generation and sound source localization did not go well.

4. Conclusion

This paper addressed 4D reconstruction by treating static and dynamic objects separately in a dynamic environment. To achieve this, we proposed Audio-Visual SfM by integrating visual SfM and sound source localization. We constructed a prototype system and validated it in a case study experiment to reconstruct a scene with the Co-Fusion dataset. The experimental results showed that audio and visual information can be disambiguated from each other for an SfM based 3D reconstruction task. Future work includes extensive evaluation of the proposed method and development of a more sophisticated audio-visual integration framework.

References

[1] R. Hachiuma, C. Pirchheim, D. Schmalstieg, and H. Saito. Detectfusion: Detecting and segmenting both known and unknown dynamic objects in real-time slam. In *BMVC*, 2019. 1

[2] H. Kaiming, G. Georgia, D. Piotr, and G. Ross. Mask r-cnn. In *ICCV*, 2017. 2

[3] 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. 2

[4] M. Rünz and L. Agapito. Co-fusion: Real-time segmentation, tracking and fusion of multiple objects. In *ICRA*, pages 4471–4478, 2017. 1, 3

[5] 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. 1

[6] R. Schmidt. Multiple emitter location and signal parameter estimation. *IEEE Transactions on Antennas and Propagation*, 34(3):276–280, 1986. 2

[7] Schönberger, J. Lutz, Frahm, and Jan-Michael. Structure-from-motion revisited. In *CVPR*, 2016. 2

[8] B. Xu, W. Li, D. Tzoumanikas, M. Bloesch, A. Davison, and S. Leutenegger. Mid-fusion: Octree-based object-level multi-instance dynamic slam. In *ICRA*, 2019. 1