Multi-Speaker Tracking from Azure Kinect: Performance Evaluation using Generalized Optimal Sub-Pattern Assignment

Muhammad Atiff Zakwan Bin Mohd Ariffin[1], Siti Nur Aisyah Binti Mohd Robi[2], Mohd Azri Bin Mohd Izhar[3], Norulhusna Binti Ahmad[4]

[1,2,3,4]Razak Faculty of Technology and Informatics, Universiti Teknologi Malaysia
[1]mazakwan2@graduate.utm.my, [2]aisyah98@graduate.utm.my, [3]mohdazri.kl@utm.my, [4]norulhusna.kl@utm.my

Article history

Received:
28 September 2023

Received in revised form:
10 Nov 2023

Accepted:
16 Nov 2023

Published online:
18 Dec 2023

*Corresponding author
mazakwan2
@graduate.utm.my

Abstract

The software and the hardware features of Azure Kinect show the potential for its application in multiple speaker tracking. We then evaluate the system's capabilities by conducting tests using our given setup, including various scenarios to evaluate its tracking performance. The tracking performance is calculated using generalized optimal sub-pattern assignment (GOSPA) and multiple objects tracking accuracy (MOTA) metrics. It has been found that the body tracking algorithm can perform well in certain multi-speaker tracking conditions.

Keywords: *Multi-speaker tracking, Azure Kinect device, optimal sub-pattern assignment, body tracking, performance evaluation.*

1. Introduction

Speaker tracking is a technique that enables users to locate the spatial position and movement of active human speech in a conversation or scene. Many speaker tracking methods can detect a single speaker only [1], [2], others can detect multiple concurrent speakers [3]–[5], and some literature can detect in a reverberant environment [6] too. These techniques need several microphones; the microphones are distributed around the scene or in the same common area, also known as a microphone array. Moreover, this technique can be supported with additional video frames and depth data captured from other devices' sensors or sensors in the same integrated device that comes with the microphone array.

Microsoft Azure Kinect is an integrated device that may help with speaker tracking activity. It has a red, green, and blue wavelengths (RGB) camera, depth sensor, infrared wavelength (IR) camera, seven microphone array, and Inertial Measurement Unit (IMU) sensor. Hence, it can record or stream all desired information: audio signals, video frames, and 3D positional data. The multi-sensing device also can sync between other similar devices. In addition, Microsoft also developed the Azure Kinect Body Tracking (AKBT) Software Development Kit

* Corresponding author. mazakwan2@graduate.utm.my

(SDK) to be used with the hardware. The AKBT SDK can simultaneously track joint locations, joint orientations, and temporal identities of multiple people in the scene. Because of its features, we decided to use this integrated device for this research.

An example of past research related to sound source localization can be found in [7], where they introduced a novel approach to locating a sound source behind an obstacle using the microphone array from the older version of Kinect. This generation of Kinect uses a four-linear microphone array. Using the newer generation with seven circular microphone arrays could improve the tracking accuracy. Meanwhile, past research on body tracking can be found in [8, 9]. From [8], they found that using Azure Kinect for gait feature extraction can perform effectively and closely similar to the golden standard, the Vicon system. Therefore, the Azure Kinect could perform exceptionally well in the speaker tracking domain. In [9], they used an OSPA-based algorithm to assess the pose quality of human performers. However, their research used a monocular camera, thus only considering the two-dimensional analysis of OSPA localization error. Expanding from our past work [10], we add another metric, the generalized optimal sub-pattern assignment (GOSPA), to evaluate the tracking performance of a new set of sequences. Compared to Multiple Object Tracking Accuracy (MOTA), which measures the percentage of correct predictions, GOSPA measures how close are the correct predictions (localization error) and how many missed and false targets (cardinality error) within a given cut-off distance.

This paper aims to investigate the tracking performance of the Azure Kinect in multiple speakers tracking scenarios. Our research utilizes the Azure Kinect for capturing and recording the audio signal, visual frames, and depth data of the speech and movement of speakers inside the sequence. Then, we evaluate the tracking performance using the GOSPA and MOTA metrics.

This research paper consists of five sections. In Section 2, we present the method and the performance metrics that we used in this research. Section 3 presents the experimental results. Then, Section 4 discusses the results found in the experiment. Finally, we conclude this research in section 5.

2. Methods

2.1. Azure Kinect Body Tracking SDK

Software developers use AKBT SDK to track bodies inside the scene. The body tracker can detect a person in the scene using its 32 joints skeleton model, as shown in Figure 1. However, only the head joint's position and identity (ID) will be used in this research. The head joint is used because it is the closest to the mouth position, which is the crucial part of speaker tracking.

Figure 1. AKBT Skeleton Model in Action

2.2. Performance Metrics

In this research, we will use two metrics to evaluate the tracking performance of the AKBT. The two metrics are GOSPA and MOTA.

2.2.1 GOSPA: GOSPA is a metric that evaluates the performance of multi-object trackers by combining the localization and cardinality errors. It is the generalized form of OSPA, which does not reward the tracker for minimizing the number of false or missed targets. From [11], the GOSPA equation is shown in Equation 1.

$$\text{GOSPA}(i) = \sum_i loc\ error + \frac{c}{2}\left(\sum_i miss + \sum_i false\right) \quad (1)$$

From the above equation, i is the i -th frame, $loc\ error$ is the Euclidean distance between the position of the tracked head joint and the actual position of the head joint, c is the cut-off distance, $miss$ is the number of missed targets, and $false$ is the number of false targets. The lower the GOSPA value, the closer the tracker is to the ground truth, and the fewer the number the tracker produces a missed or false target. The cut-off distance is 100 mm.

2.2.2. MOTA: MOTA is a metric that shows the accuracy of the object tracker. From [12], the MOTA equation is shown in Equation 2.

$$\text{MOTA} = 1 - \frac{\sum_i(\text{FP}_i + \text{FN}_i + \text{IDS}_i)}{\sum_i \text{GT}_i} \quad (2)$$

From the above equation, i is the i-th frame, FP is the number of false positive detections, FN is the number of false negative detections, IDS is the number of swapped IDs, and GT is the number of ground truths. The higher the MOTA value, the better the accuracy of the object tracker.

2.3. Azure Kinect Setup

For this research, the setup that we use for the Azure Kinect device can be seen in Table 1. However, in this research, we have yet to consider audio signals in the analysis.

Table 1: Azure Kinect Setup

Variable		Value
Colour Resolution		1080P
Colour Format		MJPG
Depth Mode		NFOV 2X2 BINNED
Framerate (FPS)		15
IMU		Disabled
Sync mode		Master / Subordinate
Firmware	**RGB camera**	1.6.110
	Depth camera	1.6.79
Microphone		7 channels, circular array
Audio properties		16kHz, 32-bit float

2.4. Dataset Capture

The total number of sequences captured is six. The number of speakers that we use is around one to three people. For verbal activity, some sequences have concurrent speech while others do not. Meanwhile, for motion activity, it is either the speakers walking around inside the field of view of the Azure Kinect or speakers (one at a time) walking outside the field of view. Table 2 summarizes the details of each sequence, while Figure 2 shows the initial frame (frame #1) in sequence 1.

Table 2: Sequence Details

Sequence		1	2	3	4	5	6
Duration (sec)		63	64	44	23	33	30
Number of speakers		3	3	3	2	2	1
Verbal Activity	**One by one**	✓	✓	✓	✓	✓	✓
	Concurrent	×	×	✓	×	✓	-
Motion	**Move within**	✓	✓	✓	✓	✓	✓

Activity	Move outside	×	✓	×	✓	✓	✓
File size (MB)	MKV	845	859	584	310	439	393
	WAV	26.0	26.6	18.1	9.76	13.7	12.6

Figure 2. Initial Frame (Frame #1) in Sequence 1

3. Results

Table 3 shows each sequence's minimum, maximum, and average GOSPA and MOTA values.

Table 3: Experimental Results

Sequence	GOSPA (mm)			MOTA (%)
	Minimum	Maximum	Average	
1	1.9959	159.9002	23.4818	98.24
2	4.8239	187.4675	70.1198	68.03
3	6.3765	179.9019	43.0784	96.27
4	1.7784	113.8740	34.9583	82.01
5	1.0849	83.7220	37.9981	78.49
6	0.2719	72.0102	22.4998	73.59

4. Discussions

According to the data presented in Table 3, it can be observed that the minimum GOSPA metrics tend to decrease in value as the number of targets being tracked decreases. The greater the number of targets, the more significant the impact on the localization error as measured by the GOSPA metric. Hence, situations involving three speakers have the highest minimum GOSPA metrics.

Sequences with more speakers exhibit a higher maximum GOSPA value due to the increased localization error that must be considered. Moreover, too many speakers may increase the number of missed or false targets, leading to a higher penalization in the GOSPA metric. Sequences with fewer speakers, including those with only one or two, show smaller maximum GOSPA values.

Overall, it can be observed that sequence 6 exhibits the lowest average GOSPA measure. This value may be attributed to the AKBT algorithm only monitoring a single target in this scenario. In the sequence where there are two speakers, it can be observed that both sequences have comparable values. However, it is notable that sequence 4 shows a better GOSPA average. In scenarios involving three speakers, it can be observed that AKBT exhibits superior tracking performance in sequence 1. In Sequence 2, the average GOSPA is the worst. As a result of its motion activity, most of the time, there will be a speaker that is outside of the view of the Azure Kinect, sample frame of a speaker about to walk out of frame can be seen in Figure 3.

Figure 3. A Speaker About to Walk Out of Frame (Frame #51) in Sequence 2

Meanwhile, for MOTA metrics, it is observed that scenarios wherein speakers walk inside the field of view throughout the sequences have higher MOTA values. These are mainly attributed to the reliable detection capabilities of the AKBT, except for instances where occlusion occurs between speakers. Example of frame

where speaker occlusion happen can be seen in Figure 4. Moreover, as the duration of the speaker's absence from the frame increases, the MOTA value decreases due to the tracker's inability to detect them visually. Consequently, it can be observed that sequences 1 and 3 exhibit the highest MOTA value since all the speakers consistently remain within the boundaries of the frame.

Figure 4. A Speaker is Being Occluded (Frame #198) in Sequence 3

5. Conclusion

In this paper, we have studied the performance of the AKBT in multi-speaker tracking. We have recorded six sequences with various numbers of speakers, i.e., one to three speakers; different verbal actions, i.e., single or concurrent speech; and different motion activities, i.e., moving inside or outside the frame of view. Performance evaluation has been measured using GOSPA and MOTA metrics. Subsequently, we found that the AKBT has better GOSPA when fewer speakers move only inside the frame of view. Also, the AKBT would have good MOTA if the speakers were visually available in the frame of view. In conclusion, AKBT is a good tracker for multi-speaker tracking if the speakers are not walking outside the frame of view and the number of speakers is small. For future works, we will include the speech signals of the speakers in our analysis.

Acknowledgements

This research was supported in part by the Ministry of Higher Education (MoHE) of Malaysia through the Fundamental Research Grant Scheme (FRGS/1/2021/TK0 /UTM/02/67) and in part by UTM SPACE Contract Research Grant UTMSPC1.16 (R.K130000.7756.4J554).

References

[1] X. Qian, A. Brutti, O. Lanz, M. Omologo, and A. Cavallaro, "Multi-Speaker Tracking from an Audio-Visual Sensing Device," *IEEE Trans Multimedia*, vol. 21, no. 10, pp. 2576–2588, Oct. 2019, doi: 10.1109/TMM.2019.2902489.

[2] X. Qian, A. Brutti, O. Lanz, M. Omologo, and A. Cavallaro, "Audio-visual tracking of concurrent speakers," *IEEE Trans Multimedia*, 2021, doi: 10.1109/TMM.2021.3061800.

[3] X. Qian, A. Brutti, M. Omologo, and A. Cavallaro, "3D audio-visual speaker tracking with an adaptive particle filter," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, Mar. 2017, pp. 2896–2900. doi: 10.1109/ICASSP.2017.7952686.

[4] F. Sanabria-Macias, M. Marron-Romera, and J. Macias-Guarasa, "3D audiovisual speaker tracking with distributed sensors configuration," in *European Signal Processing Conference*, European Signal Processing Conference, EUSIPCO, Jan. 2021, pp. 256–260. doi: 10.23919/Eusipco47968.2020.9287677.

[5] M. Sewtz, T. Bodenmuller, and R. Triebel, "Robust MUSIC-Based Sound Source Localization in Reverberant and Echoic Environments," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, Oct. 2020, pp. 2474–2480. doi: 10.1109/IROS45743.2020.9340826.

[6] J. Zhao *et al.*, "Audio-Visual Tracking of Multiple Speakers Via a PMBM Filter," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, May 2022, pp. 5068–5072. doi: 10.1109/ICASSP43922.2022.9747595.

[7] I. An, Y. Kwon, and S. Yoon, "Diffraction- and Reflection-Aware Multiple Sound Source Localization," *IEEE Transactions on Robotics*, vol. 38, no. 3, pp. 1925–1944, Jun. 2022, doi: 10.1109/TRO.2021.3118966.

[8] Y. Zeng, L. Wu, and D. Xie, "Gait Analysis based on Azure Kinect 3D Human Skeleton," in *2021 International Conference on Computer Information Science and Artificial Intelligence (CISAI)*, IEEE, Sep. 2021, pp. 1059–1062. doi: 10.1109/CISAI54367.2021.00212.

[9] Y. Zhu, S. Liang, P. Li, and X. Wu, "Robust Human Pose Quality Assessment Using Optimal Sub-Pattern Assignment," in *2022 11th International Conference on Control, Automation and Information Sciences (ICCAIS)*, IEEE, Nov. 2022, pp. 419–423. doi: 10.1109/ICCAIS56082.2022.9990031.

[10] M. A. Z. Bin Mohd Ariffin, S. N. A. B. Mohd Robi, M. A. Bin Mohd Izhar, and N. B. Ahmad, "Performance Evaluation of the Azure Kinect Body Tracking Algorithm for Multi-Speaker Tracking," in *2022 IEEE Symposium on Future Telecommunication Technologies (SOFTT)*, IEEE, Nov. 2022, pp. 67–71. doi: 10.1109/SOFTT56880.2022.10009962.

[11] A. S. Rahmathullah, A. F. Garcia-Fernandez, and L. Svensson, "Generalized optimal sub-pattern assignment metric," in *2017 20th International Conference on Information Fusion (Fusion)*, IEEE, Jul. 2017, pp. 1–8. doi: 10.23919/ICIF.2017.8009645.

[12] Y. Xu, A. sep, Y. Ban, R. Horaud, L. Leal-Taixe, and X. Alameda-Pineda, "How to Train Your Deep Multi-Object Tracker," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jun. 2020, pp. 6786–6795. doi: 10.1109/CVPR42600.2020.00682.