35th Annual International Conference of the IEEE EMBS
Osaka, Japan, 3 - 7 July, 2013

Improvement of acoustic fall detection using Kinect depth sensing

Yun Li, *IEEE student member*, Tanvi Banerjee, *IEEE student member*, Mihail Popescu, *senior IEEE member*, Marjorie Skubic, *senior IEEE member*

Abstract — The latest acoustic fall detection system (acoustic FADE) has achieved encouraging results on real-world dataset. However, the acoustic FADE device is difficult to be deployed in real environment due to its large size. In addition, the estimation accuracy of sound source localization (SSL) and direction of arrival (DOA) becomes much lower in multi-interference environment, which will potentially result in the distortion of the source signal using beamforming (BF). Microsoft Kinect is used in this paper to address these issues by measuring source position using the depth sensor. We employ robust minimum variance distortionless response (MVDR) adaptive BF (ABF) to take advantage of well-estimated source position for acoustic FADE. A significant reduction of false alarms and improvement of detection rate are both achieved using the proposed fusion strategy on real-world data.

I. Introduction

Falls have become increasingly concerning among older adults. A recent report from CDC [1] shows that one third of older adults in US fall each year. The cost for the direct medical care of the fall-related health problems has reached to $30 billion [1]. The delayed intervention and unreported falls put that population at higher risk of early death [2]. These issues have motivated a variety of research in fall detection in the past several years.

The latest version of acoustic FADE proposed in [3] has achieved encouraging performance both in laboratory and real apartment settings in which background noise and audio interference are considered. Although the improvement of acoustic FADE in challenging acoustic environments could be attributed to its 2-dimensional circular array with higher sound source localization accuracy, we believe that it is mostly due to its large array dimension (0.5m diameter) that is difficult to install and conceal in home environment. In addition, acoustic FADE is most likely to fail in locating the source of a fall using SSL or DOA estimator when multiple interferences, such as phone ringing and TV noise, are presented. In multi-interference acoustic environments, tasks such as source identification and separation require complex and time-consuming techniques [4-6]. To deal with multi-interference issues in practice, obtaining source position information from another type of sensor can greatly improve fall detection performance.

In this paper, we use for fall detection a Microsoft Kinect device [7], which consists of two main types of sensors: a 4-element linear microphone array and a depth sensor. Since the depth sensor facilitates human body detection, Kinect has been largely investigated for activity recognition tasks [8-10].

To take advantage of superior source localization provided by Kinect in comparison to a conventional delay-and-sum (DSBF) approach, we use MVDR-ABF. MVDR-ABF results in higher selectivity and quality of sound from the direction of interest with minimum power of interferences-plus-noise signals (IPNs) from other directions [11-12]. In this paper, we propose a fusion strategy by utilizing source position information from depth sensing to help obtain higher reliability in both height discrimination and beamforming of falls [3]. The data was collected at TigerPlace, an independent living facility for older adults in Columbia, Missouri. For two years now, we have been collecting Kinect depth data in ten TigerPlace apartments to capture the activities of daily living of the residents [13-14]. The experimental results show that Kinect-based acoustic FADE performance is significantly improved by utilizing the proposed fusion strategy.

This paper is structured as follows: in section II we present the methodology of our study, in section III we discuss the results and in section IV we provide conclusions and on-going work.

II. Methodology Used for Kinect FADE

A. Centre of Mass (COM) extraction using depth sensing

The values returned by the depth sensor are first converted into 3D point clouds by estimating the intrinsic and extrinsic parameters associated with the Kinect. Adaptive background estimation helps filtering out objects that are moved around in a continuously changing environment such as an inhabited apartment. Once a foreground object is extracted, its centroid information is tracked and stored until it either blends into background or goes outside the field of view. Hence, the centroid information or 3D COM position of each foreground object is tracked separately and stored [13].

B. Description of robust MVDR adaptive beamforming

1) Determination of DOA from COM position

The 3D Cartesian coordinate system of Kinect has its origin position $\boldsymbol{o}$ at the CMOS image sensor [7] which produces depth images. Under far-field assumption, plane

Y.L. and T.B. are with Elect. and Comp. Eng. Dept., University of Missouri, Columbia, MO 65211, USA (corresponding author's e-mail: yl874@mail.missouri.edu).

M.P. is with the Health Mngmt. & Inform. Department, University of Missouri, Columbia, MO 65211, USA (email: popescum@missouri.edu).

M.S is with Elect. and Comp. Eng. Dept., University of Missouri, Columbia, MO 65211, USA (email: SkubicM@missouri.edu).

sound waves propagate to the microphones with the same incident angle θ as shown in Fig. 1, DOA can be estimated as $\hat{\theta}$ computed by the following equation:

$$\hat{\theta} = \tan^{-1}\left(p_z/\sqrt{p_x^2 + p_z^2}\right), \quad (1)$$

in which p_x, p_z are x and z coordinate values of COM position $\boldsymbol{p}$. Equation (1) estimates DOA as the azimuthal angle between origin-to-COM vector and z-axis vector by ignoring the elevation value p_y due to the symmetric property of beam pattern of a linear microphone array.

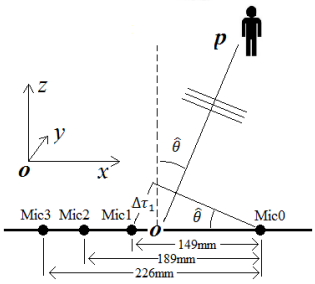

Fig. 1 Kinect microphone array configuration in far-field acoustic model.

2) Computation of the steering vector of Kinect microphone array

As depicted in Fig. 1, once we know $\hat{\theta}$, the delay of each microphone with respect to Mic0, $\Delta\tau_k, k = 1,2,3$ is calculated by

$$\Delta\tau_k(\hat{\theta}) = d_{k0} \sin\hat{\theta}/c \quad (2)$$

where d_{k0} represents the distance between Mic$k$ and Mic0, c is the locally measured sound propagation velocity. Thus, we can write a delay vector $\Delta\boldsymbol{\tau}$ given $\hat{\theta}$ as

$$\Delta\boldsymbol{\tau}(\hat{\theta}) = \left[0\ \Delta\tau_1(\hat{\theta})\ \Delta\tau_2(\hat{\theta})\ \Delta\tau_3(\hat{\theta})\ \right]^T. \quad (3)$$

Therefore, the steering vector $\boldsymbol{s}$ with respect to a specific frequency bin ω given $\hat{\theta}$ is defined as

$$\boldsymbol{s}(\omega;\hat{\theta}) = \exp\left(-\mathrm{j}\omega\Delta\boldsymbol{\tau}(\hat{\theta})\right). \quad (4)$$

3) Description of MVDR ABF

MVDR is a widely used adaptive beamforming technique that determines the weights of the sensor outputs by a constrained minimization of the interference and background noise power. The constraint is that the gain in the direction steered to the source of interest is unity. Alternatively, as long as DOA is estimated correctly, the signal of interest travels through MVDR without any distortion while signals from other directions are suppressed. Suppose $\boldsymbol{h}(\omega;\hat{\theta})$ is a column vector of weights need to be estimated given the DOA, $\hat{\theta}$. $\boldsymbol{x}(\omega)$ is a column vector of sensor output spectrum. The beamformed output $y(\omega)$ is defined as (for simplicity, we ignore 'ω' and '$\hat{\theta}$' in these terms such that $\boldsymbol{h} = \boldsymbol{h}(\omega;\hat{\theta})$, etc.)

$$y = \boldsymbol{h}^*\boldsymbol{x} \quad (5)$$

where '*' represents complex conjugated transformation. The goal of MVDR can be expressed as

$$\min_{\boldsymbol{h}} E(|y|)^2 = \min_{\boldsymbol{h}} E(\boldsymbol{h}^*\boldsymbol{x}\boldsymbol{x}^*\boldsymbol{h}) \quad \text{subject to } \boldsymbol{h}^*\boldsymbol{s} = 1. \quad (6)$$

We assume that

$$\boldsymbol{x} = \boldsymbol{s}^* g + \boldsymbol{n} \quad (7)$$

where g and $\boldsymbol{n}$ represent the signal of interest and IPNs in frequency domain. If g and $\boldsymbol{n}$ are independent and uncorrelated, with the constraint $\boldsymbol{h}^*\boldsymbol{s} = 1$, one can verify that

$$\min_{\boldsymbol{h}} E(\boldsymbol{h}^*\boldsymbol{x}\boldsymbol{x}^*\boldsymbol{h}) \equiv \min_{\boldsymbol{h}} E(\boldsymbol{h}^*\boldsymbol{n}\boldsymbol{n}^*\boldsymbol{h}) = \min_{\boldsymbol{h}} \boldsymbol{h}^*\boldsymbol{Q}\boldsymbol{h} \quad (8)$$

where $\boldsymbol{Q} = E(\boldsymbol{n}\boldsymbol{n}^*)$ is the cross-spectral matrix of IPNs. The solution to the constrained optimization problem is

$$\hat{\boldsymbol{h}} = \frac{\boldsymbol{Q}^{-1}\boldsymbol{s}}{\boldsymbol{s}^*\boldsymbol{Q}^{-1}\boldsymbol{s}}. \quad (9)$$

The exact estimation of $\boldsymbol{Q}$ is unavailable, however $\boldsymbol{Q}$ at i^{th} frame can be estimated by the sample covariance matrix of the last N frames of IPNs, as expressed by

$$\hat{\boldsymbol{Q}}_i = \frac{1}{N}\sum_{k=0}^{N-1} \boldsymbol{n}_{i-k}\boldsymbol{n}_{i-k}^*. \quad (10)$$

4) Efficient and adaptive updating of inversed cross-spectral matrix $\boldsymbol{Q}^{-1}$

A broadband MVDR-ABF requires the computation of $\hat{\boldsymbol{h}}$ for each frequency bin which results to high computational cost of inversing $\boldsymbol{Q}$ for each frequency bin. In addition, the updating of $\boldsymbol{Q}$ should be adaptive to highly non-stationary sources such as speech and music. Therefore the equally weighted estimation of $\boldsymbol{Q}$ in equation (10) is not appropriate in this case. To address both problems, sequential regression (SER) as suggested in [12] is used for our case. The updating of inversed $\hat{\boldsymbol{Q}}$ is derived as the following recursive equations.

$$\begin{cases} \hat{\boldsymbol{Q}}_{i+1}^{-1} = \frac{1}{a}\hat{\boldsymbol{Q}}_i^{-1} - \left(\frac{1-a}{a}\right)\left[\frac{\hat{\boldsymbol{Q}}_i^{-1}\boldsymbol{n}_{i+1}\boldsymbol{n}_{i+1}^*\hat{\boldsymbol{Q}}_i^{-1}}{a+(1-a)\boldsymbol{n}_{i+1}^*\hat{\boldsymbol{Q}}_i^{-1}\boldsymbol{n}_{i+1}}\right] \\ \hat{\boldsymbol{Q}}_0^{-1} = (1/e)\boldsymbol{I} \end{cases} \quad (11)$$

in which a is the forgetting factor between 0 and 1, e is a small positive number and $\boldsymbol{I}$ is a 4×4 identity matrix. a determines how much information of previous $\hat{\boldsymbol{Q}}$ remains for updating new $\hat{\boldsymbol{Q}}$ by the following equation

$$\hat{\boldsymbol{Q}}_{i+1} = a\hat{\boldsymbol{Q}}_i + (1-a)\boldsymbol{n}_{i+1}\boldsymbol{n}_{i+1}^*. \quad (12)$$

Thus, once a new frame of IPNs is received, equation (11) is used to update $\hat{\boldsymbol{Q}}^{-1}$ with higher computational efficiency and better adaptation to the data. The processing prototype of Kinect FADE in the next section will describe how to detect frames of IPNs using depth sensing information.

C. Processing prototype of Kinect FADE

The flowing diagram in Fig. 2 depicts the proposed processing prototype for Kinect FADE.

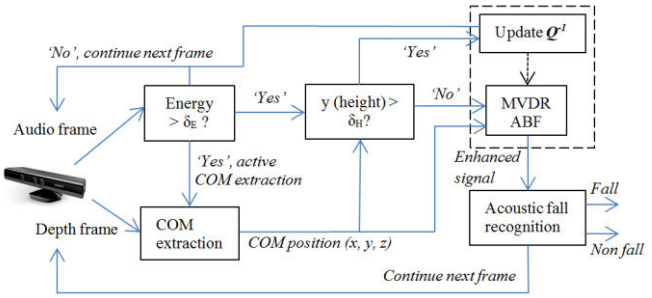

Fig. 2 Proposed processing prototype of Kinect FADE.

As shown in Fig. 2, the current audio frame will firstly pass to the activity detector using energy thresholding. Once we detect an activity, the 3D COM position of the human subject will be extracted from the current depth frame. In the height

discriminator, if the height of the subject or y dimension of COM position is less than a pre-defined threshold δ_H, the current audio frame is then passed to MVDR ABF using the most updated $\boldsymbol{Q}^{-1}$; otherwise, we consider the audio frame as IPNs and use it to update $\boldsymbol{Q}^{-1}$. The beamformed output signal is then used for fall recognition. The processing algorithms of acoustic fall recognition are not relevant for this paper but interested readers are referred to [3].

D. *Synchronization of sensors for data acquisition*

To synchronize both audio and depth sensors during data acquisition, we set the frame rate of both audio and depth sensors to 8 frames per second. Thus, a microphone reads in a new segment of 125ms audio data at each time a depth frame is read. Each new segment of audio data is combined with its previous one to form a 250ms frame for processing.

E. *Experimental data description*

The experimental data is recorded using Kinect sensors in seven TigerPlace apartments [15]. The Kinect device is placed above the apartment door viewing the entire space of the living room. The falls are performed by one female stunt actor in each apartment under the instructions of a nursing staff [16]. The stunt actor falls on a mattress placed on the floor in different parts of the room. The complete dataset DAT I and related information is presented in Table. I.

TABLE I. STUNT ACTOR FALLS OF DAT I

Room and file (audio+depth) #	# of falls	Noise and interference types, SNR or SIR when a fall occurs
R100, File1	4	Background noises, 6dB
R100, File2	4	Background noises, 6dB
R100, File3	4	Background noises, 7dB
R100, File4	4	Background noises, 8dB
R100, File5	4	Background noises, 5dB
R102, File6	4	Background noises, 6dB
R106, File7	4	Background noises, 6dB
R109, File8	4	Background noises, 5dB
R110, File9	4	Background noise +TV speech, 2dB
R111, File10	4	Background noise +TV music, 2dB
R112, File11	4	Background noises, 8dB

The duration of each file in DAT I is about 11 minutes. In all the files, background noises such as machine and fan noise are present with a signal-to-noise ratio (SNR) of around 6dB. Except for File9 and File10, all files are recorded in the acoustic environment without any interference overlapped with falls. The falls in both File9 and File10 are performed in multi-interference acoustic environment in which a loud TV audio signal containing mainly speech and music with signal-to-interference ratio (SIR) of 2dB is present for the entire duration.

We use receiver operation characteristic (ROC) curves to evaluate the fall recognition performance using another dataset (DAT II). DAT II consists of 60 fall files and 310 non fall files extracted from DAT I. Each fall and non fall file has duration of 500ms. The audio and depth frames are well synchronized for each file.

III. RESULTS AND DISCUSSIONS

A. *Improvement of FADE performance using height information from depth sensing*

To better track and analyze the motion of human subjects, foreground images which consist of silhouettes of the moving subjects are extracted (as described in Section II.A). A sequence of a typical human fall is showed in the following selected frames of foreground images.

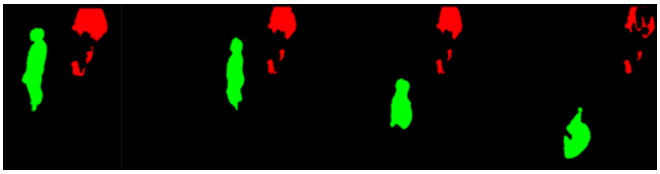

Fig. 3. Four images from a stunt actor fall. The stunt actor falls in front of a resident in the apartment. The system is able to identify both the persons in the scene (resident shown in red, stunt actor shown in green).

In Fig. 3, the two colors green and red indicate that the system is able to distinguish two or more subjects when they are far enough apart. This demonstration shows that the system is able to track the stunt actor in challenging environments when more than one person is present. The system maintains a separate track history for the stunt actor as well as the resident.

To show the advantage of the tracking ability for height discrimination in acoustic FADE (as stated in Fig. 2), the audio data from File1 is processed for detecting falls. In this data, the scenario that a stunt actor falls in front of one or more residents is included. The FADE results of using and not using height information are shown in Fig 4.

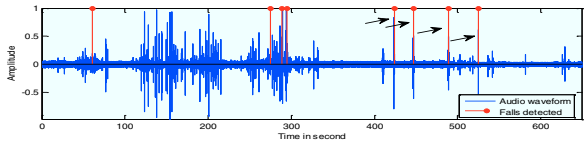

Fig. 4 (a) Detected falls with no height information (true falls are pointed at).

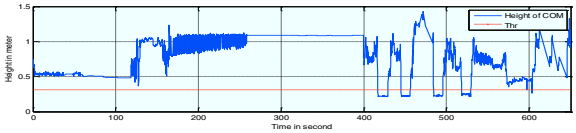

Fig. 4 (b) Height information from COM with a pre-defined threshold 'Thr'.

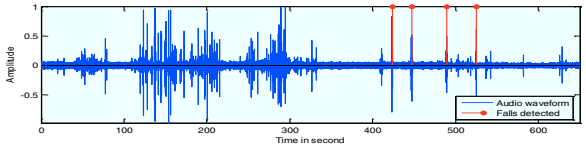

Fig. 4 (c) Detected falls with height information.

Note that in Fig. 4(c) the four false alarms from Fig. 4(a) (red markers, not marked by arrows) are removed if the discriminator threshold is selected properly as shown in Fig. 4(b). The threshold is statistically determined by first building a normal distribution on training heights then selecting the one whose probability (p-value) equals to a significance level.

B. *Fall sound enhancement using MVDR ABF*

To demonstrate that MVDR-ABF is able to enhance the fall signal by suppressing IPNs, we select a 500ms segment of audio data from File9. In this segment, a fall signal comes from an estimated DOA of 50 degrees mixed with a TV signal that comes from a fixed pre-measured DOA of -10 degrees. A previous 15-second segment of TV speech is only used as training data for updating $\boldsymbol{Q}^{-1}$ (a =0.8, e =10^{-8}). For comparison purpose, we also apply DSBF to the mixture signals. Fig. 5 shows the waveform of mixture signal from Mic1 and the enhanced fall signals obtained by both MVDR-ABF and DSBF.

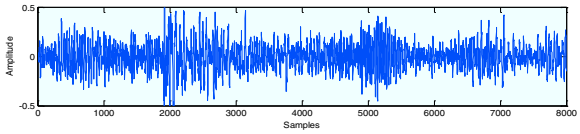

Fig. 5 (a) The mixture waveform of a fall signal and a TV speech signal.

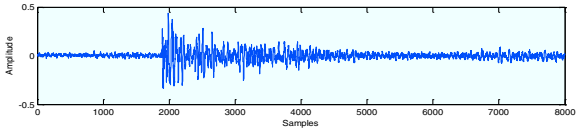

Fig. 5 (b) The enhanced fall signal using MVDR-ABF.

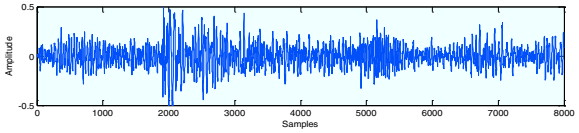

Fig. 5 (c) The enhanced fall signal using DSBF.

By inspecting above figures we see that the fall signal processed using MVDR-ABF (Fig. 5(b)) looks cleaner than the one obtained by DSBF (Fig. 5(c)).

C. *Cross-validation evaluation of FADE on DAT II*

To compare FADE performances of using and not using depth sensing, we considered three evaluation methods.
-*Method 1*: evaluation using a single microphone where Mic2 is used for training and Mic1, 3 and 4 are used for testing.
-*Method 2*: evaluation not using depth sensing, which consists in using conventional SSL and DSBF [3] (height discrimination is not considered due to bad estimation of height using a linear microphone array).
-*Method 3*: evaluation using depth sensing as proposed in this paper. The 5-fold cross-validation ROC curves of the three evaluation methods with corresponding area under the ROC curve (AUROC) are shown in Fig. 6.

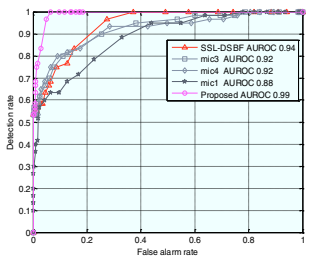

Fig. 6 Comparison of FADE performance using 3 evaluation methods.

Fig. 6 shows that by utilizing the position information from depth sensing, false alarm rate at 100% detection is reduced by about 80% (from 38% to 6.5%) as compared to the case when no depth sensing is used.

IV. Conclusion

In this paper we present a sensor fusion strategy in which tracking information obtained from Kinect depth sensing is used to improve acoustic FADE performance. We believe that the Kinect depth sensing is robust enough to provide accurate position information of the human subject of interest in multi-interference environment. In our experiments, the use of Kinect depth information for fall detection reduced the false alarm rate by about 80%.

Future work will focus on improving the Kinect depth-sound fusion strategy which fuses FADE results from both depth and audio sensing to achieve better performance.

Acknowledgements

This work has been supported in part by the NSF grant CNS-0931607.

References

[1] Center for Disease Control (2012), http://www.cdc.gov/HomeandRecreationalSafety/Falls/adultfalls.html.

[2] R. J. Gurley, N. Lum, M. Sande, B. Lo, and M. H. Katz, "Persons found in their homes helpless or dead," *N. Engl. J. Med.*, vol. 334, no. 26, pp. 1710–1716, 1996.

[3] Y. Li, K.C. Ho, and M. Popescu, "A microphone array system for automatic fall detection," *IEEE Trans. Biomedical Engineering*, vol. 59, pp. 1291-1301, 2012.

[4] J. M. Valin, F. Michaud, and J. Rouat, "Robust localization and tracking of simultaneous moving sound sources using beamforming and particle filtering," *Journal of robotics and autonomous systems,* vol. 55, pp.216-228, 2007.

[5] S. Kagami, Y. Tamai, H. Mizoguchi, and T. Kanade, "Microphone array for 2D sound localization and capture," in *Proc. Int. IEEE Robotics and Automation Conf.,* 2004, pp.703-708.

[6] G. J. Jang and T. W. Lee, "A maximum likelihood approach to single channel source separation," *Journal of Machine Learning Research*, vol. 4, pp. 1365-1392, 2003.

[7] Microsoft Corp., Redmond WA. Kinect for Xbox 360.

[8] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake, "Real-time human pose recognition in parts from a single depth image," in *IEEE International Conference on Computer Vision and Pattern Recognition*, 2011, pp. 1297-1304.

[9] E. Machida, M. Cao, T. Murao, and H. Hashimoto, "Human motion tracking of mobile robot with kinect 3D sensor," in *SICE annual Conf.*, Akita, Japan, 2012, pp. 2207-2211.

[10] I. Oikonomidis, N. Kyriazis, and A. A. Argyros, "Efficient model-based 3D tracking of hand articulations using Kinect," in *BMVC*, Dundee, UK, Aug. 2011.

[11] H. Van Trees, *Optimum Array Processing*. New York: Wiley-Interscience, 2002.

[12] P. Sinha, A. D. George, and K. Kim, "Parallel algorithm for robust broadband MVDR beamforming," *Journal of Computational Acoustics*, vol. 10, no.1, pp. 69-96, 2002.

[13] E.E. Stone and M. Skubic, "Capturing Habitual, In-Home Gait Parameter Trends Using an Inexpensive Depth Camera," in *Proc. 34th Int. IEEE EMBS Conf.*, San Diego, CA, 2012, pp. 5106-5109.

[14] T. Banerjee, J. M. Keller, and M. Skubic, "Resident Identification Using Kinect Depth Image Data and Fuzzy Clustering Techniques," in *Proc. 34th Int. IEEE EMBS Conf.*, San Diego, CA, 2012, pp. 5102-5105.

[15] The Institutional Review Board at University of Missouri has approved the research work.

[16] M. Rantz, M. Aud, G. Alexander, B. Wakefield, M. Skubic, R.H. Luke, D. Anderson, and J. Keller, "Falls, technology, and stunt actors: new approaches to fall detection and fall risk assessment," *Journal of Nursing Care Quality*, vol. 23(3), pp. 195-201, 2008.