VOL. 11, NO. 14, JULY 2016 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

AUDIO VISUAL TRACKING OF A SPEAKER BASED ON FFT AND KALMAN FILTER

Muhammad Muzammel, Mohd Zuki Yusoff, Mohamad Naufal Mohamad Saad and Aamir Saeed Malik
Centre for Intelligent Signal and Imaging Research, Electrical and Electronic Engineering Department, Universiti Teknologi Petronas, Malaysia
E-Mail: muhammad_muzammel@yahoo.com

ABSTRACT

In this paper a simple audio visual information based speaker tracking technique is proposed for indoor environment. Specifically, a Kalman filter based image processing technique is used to extract visual information, and Fast Fourier Transform (FFT) based approach is used to extract audio information for speaker tracking. Finally, a decision tree has been used to estimate the location of the speaker based on audio and visual information. One of the main advantages of the proposed technique is the use of a built-in microphone of the tracking camera; which makes this technique cost effective and simple. We have examined our method with case studies from the online SPEVI database. The proposed technique shows the best detection and works properly even when the speaker is not visible.

Keywords: speaker tracking, kalman filter, image processing, FFT, audio visual tracking.

INTRODUCTION

A speaker tracking for indoor environments has received much interest in the fields of computer vision and signal processing in the past decades. Speaker tracking may be achieved in a single modality domain through video (Winkler, Michael and Mühlhäuser, 2014) or audio (Cobos, Lopez and Martinez, 2011).

For visual tracking, a description of the object is required for tracking through the camera. The description can be the template image of object, a shape, texture, color model or something alike. Recently, gradient features have been proved advantageous in speaker detection (Pang *et al*., 2011), (Sabzmeydani and Mori, 2007). To increase the discriminative power, color descriptors have been proposed for speaker detection (Khan, *et al*., 2012), (Kviatkovsky, Adam, and Rivlin, 2013), (Van, Gevers and Bagdanov, 2006). Color features represent the global information of images, which are relatively independent from the viewing angle, translation, and rotation of the objects and regions of interest. Texture features can also be used for speaker tracking (Kellokumpu Zhao, and Pietikäinen, 2011), (Shotton *et al*., 2009). Some other researchers (Wang *et al*., 2009), (Xia and Aggarwal, 2013) also proposed spatiotemporal shifts for speaker tracking for indoor environment.

It might be possible that a single visual feature may not give us a desired accuracy. For example, the objects with the same color histogram can be completely different in texture; thus color histogram cannot provide enough information for speaker tracking. Therefore these features can be used together to achieve better performance (Shotton, Blake and Cipolla, 2008).

However, video tracking is affected by the limited field of view, occlusions, and changes in appearance and illumination. Next, it is a crucial and hard task to build an initial object description because the quality of the description directly relates to the quality of the tracking process.

On the other hand, audio tracking is not restricted by these limitations and it can be detected even when the speaker is hidden from the camera while emitting sound (Cobos, Lopez and Martinez, 2011), (Lathoud and Magimai-Doss, 2005). Mostly for an audio tracking, multiple microphones are used to detect the location of a speaker (Brutti and Nesta, 2013), (Plinge and Fink, 2014), (Plinge and Fink, 2014).

The main challenge for audio tracking techniques is that a target can be silent in some periods of time and not detectable by audio measurements. Other than that, these tracking techniques are prone to the errors caused by acoustic noise, room reverberations and the intermittency between utterance and silence (Kilic *et al*., 2013). Mostly for these tracking techniques, synchronization of multiple microphones is required to estimate the speaker position. Synchronization of multiple microphones increases the complexity of the audio tracking algorithms.

Many researchers (Blauth *et al*., 2012), (Hoseinnezhad *et al*., 2011), (Kilic *et al*., 2013) proposed speaker detection based on both audio and visual information. These techniques also used multiple microphones to estimate the position, when the speaker is invisible at the camera. Therefore for the above techniques, synchronization of multiple tracking microphones and camera is required to estimate the speaker position. Similarly combining audio and visual tracking makes the algorithm more complex.

In brief, finding the location of a speaker by a single microphone is very challenging. In this paper a simple audio visual tracking technique for a speaker in an indoor environment has been proposed.

DATA SET

The Motinas Room105 online data (EP/D033772/1, http:// www.eecs.qmul.ac.uk / ~andrea / spevi.html) are used for tracking the person. The dataset consists of a recording of a person in a room with a video camera and two microphones.

The dataset was recorded in a room with reverberations. The experiment instrument setup primarily consists of: 1) Image size with 360 x 288 pixels; 2) images

format with 8 bit color AVI; 3) 25fps video sampling rate; and 4) audio sampling rate is 44.1 kHz.

PROPOSED TECHNIQUE

The proposed audio visual technique is based on a Kalman filter for video tracking, and an FFT based approach is used for an audio tracking. A decision tree has been proposed to eventually combine both audio and visual tracking. The uniqueness of this proposed tracking technique is the use of internal microphone of a tracking camera. Therefore for proposed tracking technique, there is no need to install arrays of microphones for audio detection. The proposed technique has been simulated in MATLAB R2014a.

Visual tracking

Kalman filtering is used for the visual information extraction of a speaker. In a given technique, a Kalman filter contestant acceleration model is implemented. The Kalman filter was proposed in 1960 for the use in optimal control of navigation systems based on non-imaging information. Afterwards, the Kalman filter has also been widely used in image processing area since the early year of 1970s.

Kalman filter based object detection and object location prediction from visual information is more accurate when the object is still or moving with steady speed. This filter also provides convincing results for object detection and object location prediction even when there is a constant change in object speed. For a random movement or random variation in object speed, the Kalman filter gives reliable visual detection information. But the results will be less convincing and less accurate when an object is out from the camera capturing area or completely hides behind some other object.

In most of the cases, the movement of the speaker is random or there is a random variation in a speaker speed. Even for the given online data set, the speed of the speaker is varying randomly. Therefore in the proposed technique, only Kalman filter visual detection information is used. For scenario when the speaker moves away from the camera capture area or completely hides behind some object; an audio tracking has been proposed to detect the speaker.

Audio tracking

For audio tracking in the proposed technique, initially audio data have been collected from the internal microphone of a tracking camera. Then audio data were sampled at 44.1 kHz and pre-filtered initially in order to reduce the reverberation effect. The proposed technique for audio tracking is shown in Figure-1.

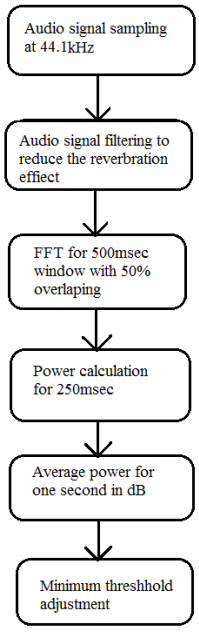

Figure-1. Proposed technique for audio tracking.

After the pre-filtration, the dataset has been divided into 500msec window and Fast Fourier Transform (FFT) has been applied on it with 50% overlapping. The FFT is a faster version of the Discrete Fourier Transform (DFT). The FFT utilizes algorithms to do the same tasks as the DFT, but in much less time.

Afterward, the power of a FFT signal is calculated for 250msec duration and then the average power has been computed for one second. From the dataset, it has been observed that the speaker keeps quit for some duration; therefore a threshold level has been adjusted for the speaker accurate location detection.

Audio visual tracking

Lastly, both the audio and visual tracking techniques have been combined by using a decision tree. The decision tree is a simple and fastest method to make a selection. The proposed decision tree is shown in Figure 2.

As when the speaker is in the camera detection area, the Kalman filter provides more accurate visual information; therefore the decision tree relies on Kalman filter image information. On the other hand, when the speaker is out of the camera capture zone due to the factor of the non-linear movement of the speaker, the decision tree used the audio signal power to calculate speaker position.

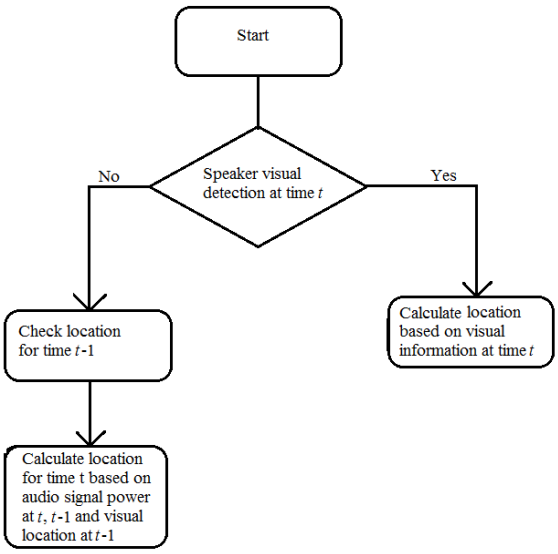

Figure-2. Decision tree for audio visual tracking.

Mathematically, t is assumed as the time instant for which the speaker is not visible at the camera. The speaker was visible at t-1. The power at time t and t-1 are $P(t)$ and $P(t\text{-}1)$, respectively. Let the position of the speaker at instant t-1 be (X_2, Y_2). The position of a speaker at time t (X_1, Y_1) can be calculated as.

$$X_1 = (X_2 + (P(t) - P(t\text{-}1)) / C_1) \quad (1)$$

$$Y_1 = (Y_2 + (P(t) - P(t\text{-}1)) / C_2) \quad (2)$$

Where C_1 and C_2 are constants and are used to adjust the amplitude values for coordinates.

For time t+1, if the speaker is visible then visual information is used to detect the position. However if the speaker is not visible, the calculated position at time t is used to find the position at time t+1.

SIMULATION RESULTS

We have examined the ability of our method to track a speaker in the audio-visual sequence from the SPEVI database. For the proposed technique, accuracy has been computed based on the ground truth that was provided with the online dataset. The result section has been divided into three parts namely visual, audio and audio visual tracking.

Visual tracking

For the proposed technique, a Kalman filter has been used for visual tracking of the speaker. In a given data set, there is a random variation in a speaker speed. Therefore, the Kalman filter gives admirable results only when the speaker is visible in the camera. Figures 3 and 4 show the snapshots of the video tracking results in the sequence.

Figure-3. Snapshot of speaker tracking using Kalman filter (frame-21).

Figure-4. Snapshot of speaker tracking using Kalman filter (frame-312).

Audio tracking

The original audio signal and the filtered signal plots are shown in Figure-5.

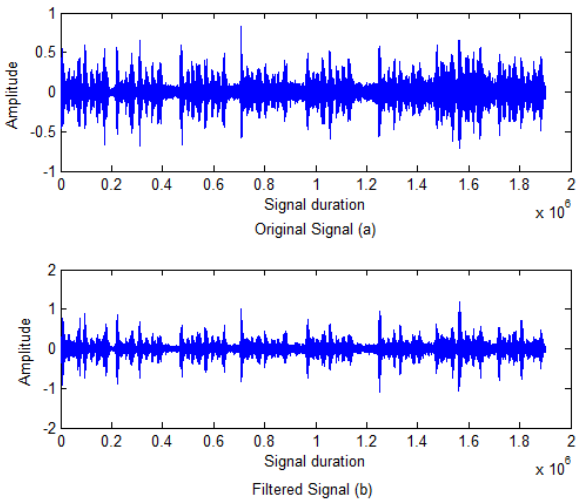

Figure-5. Original and filtered audio signal plots in Matlab.

Due to the random variation in a speaker speed, for audio tracking the power features have been computed.

After the proposed technique was applied, the power feature has been obtained as shown in Figure-6.

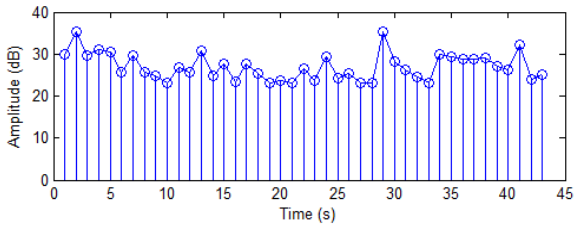

Figure-6. Power feature for audio tracking.

Audio visual tracking

Finally, a decision tree has been proposed to find the speaker position based on the audio and visual tracking. The proposed technique gives a good result when the person is either visible or invisible. Figures 7 and 8 show the snapshots of the audio-video tracking results in the sequence.

Figure-7. Snapshot of speaker tracking (frame250).

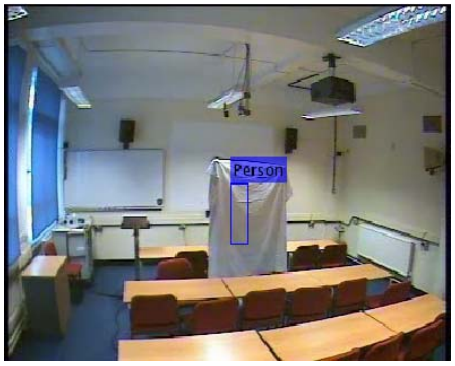

Figure-8. Snapshot of speaker tracking when he is invisible (frame265).

Table-1 shows the results produced in the present work. For a given technique, accuracy, error rate and average computation time has been calculated. The accuracy has been computed based on the ground truth provided with the online dataset. The accuracy achieved for the proposed technique is 95%.

Referring to Table-1, it was observed that proposed technique gives comparable results with literature reported (Hoseinnezhad *et al*., 2011), (Zhou, Taj, and Cavallaro, 2007). For comparison study, sequence-2 results of (Hoseinnezhad *et al*., 2011) have been used. Both (Hoseinnezhad *et al*., 2011) and (Zhou, Taj, and Cavallaro, 2007) used multiple microphones for audio processing. While the proposed technique used internal microphone of camera, which make it much simpler as compared to others.

Table-1. Results of proposed technique.

Reference	**Accuracy** (%)	**Error rate** (%)	**Average computation time** (msec)
This work	95	5.02	18.2
Zhou, Taj, and Cavallaro. (2007)	NA	5.2	NA
(Hoseinnezhad *et al*., 2011)	95	NA	NA

CONCLUSIONS

From this research, it has been concluded that the proposed technique is cost effective and simple, due the use of a built-in microphone of a tracking camera. As a built-in microphone is used in the proposed technique, therefore the synchronization of multiple microphones is not required. The reverberation effect in audio signal has been minimized by pre-filtration and its increased the efficiency of tracking. The proposed audio visual tracking technique is very simple and easily implementable. Lastly, a decision tree shows efficient performance in audio and visual tracking.

ACKNOWLEDGEMENT

We express gratitude and acknowledge to the Department of Electronic Engineering - Queen Mary, University of London, for publically making available the dataset of MOTINAS project (EP/D033772/1). This research is supported by the Centre for Graduate Studies (CGS), Universiti Teknologi PETRONAS, Malaysia;

through the grant of a Graduate Assistantship (GA) scholarship.

REFERENCES

Blauth, D. A., Minotto, V. P., Jung, C. R., Lee, B. and Kalker, T. 2012. Voice activity detection and speaker localization using audiovisual cues. Pattern Recognition Letters, 33(4), pp. 373-380.

Brutti, A. and Nesta, F. 2013. Tracking of multidimensional TDOA for multiple sources with distributed microphone pairs. Computer Speech and Language, 27(3), pp. 660-682.

Cobos, M., Lopez, J. J. and Martinez, D. (2011). Two-microphone multi-speaker localization based on a Laplacian mixture model. Digital Signal Processing, 21(1), pp. 66--76.

Hoseinnezhad, R., Vo, B. N., Vo, B. T. and Suter, D. 2011. Bayesian integration of audio and visual information for multi-target tracking using a CB-MeMBer filter. In Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference, pp. 2300--2303.

Kellokumpu, V., Zhao, G. and Pietikäinen, M. 2011. Recognition of human actions using texture descriptors. Machine Vision and Applications, 22(5), pp. 767-780.

Khan, R., Hanbury, A., Stöttinger, J. and Bais, A. 2012. Color based skin classification. Pattern Recognition Letters, 33(2), pp. 157-163.

Kilic, V., Barnard, M., Wang, W. and Kittler, J. 2013. Audio constrained particle filter based visual tracking. In Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference, pp. 3627-3631.

Kviatkovsky, I., Adam, A. and Rivlin, E. 2013. Color invariants for person reidentification. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 35(7), pp. 1622-1634.

Lathoud, G., and Magimai-Doss, M. 2005. A sector-based, frequency-domain approach to detection and localization of multiple speakers. In Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference, Vol. 3, pp. iii-265.

Pang, Y., Yuan, Y., Li, X. and Pan, J. 2011. Efficient HOG human detection. Signal Processing, 91(4), pp. 773-781.

Plinge, A. and Fink, G. 2014. Geometry calibration of multiple microphone arrays in highly reverberant environments. In Acoustic Signal Enhancement (IWAENC), 2014 14th IEEE International Workshop, pp. 243-247.

Plinge, A. and Fink, G. 2014. Multi-speaker tracking using multiple distributed microphone arrays. In Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference, pp. 614-618.

Sabzmeydani, P. and Mori, G. 2007. Detecting pedestrians by learning shapelet features. In Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference, pp. 1-8.

Shotton, J., Blake, A. and Cipolla, R. 2008. Efficiently Combining Contour and Texture Cues for Object Recognition. In BMVC, pp. 1-10.

Shotton, J., Winn, J., Rother, C. and Criminisi, A. 2009. Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context. International Journal of Computer Vision, 81(1), pp. 2-23.

Van de Weijer, J., Gevers, T. and Bagdanov, A. D. 2006. Boosting color saliency in image feature detection. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 28(1), pp. 150-156.

Wang, H., Ullah, M. M., Klaser, A., Laptev, I. and Schmid, C. 2009. Evaluation of local spatio-temporal features for action recognition. In BMVC 2009-British Machine Vision Conference, BMVA Press, pp. 124.1--124.11.

Winkler, M., Michael Höver, K., and Mühlhäuser, M. 2014. A depth camera based approach for automatic control of video cameras in lecture halls. Interactive Technology and Smart Education, 11(3), pp. 169-183.

Xia, L. and Aggarwal, J. K. 2013. Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera. In Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference, pp. 2834-2841.

Zhou, H., Taj, M., and Cavallaro, A. 2007. Audiovisual tracking using STAC sensors. In Distributed Smart Cameras, 2007. ICDSC'07. First ACM/IEEE International Conference, pp. 170-177.