I.J. Image, Graphics and Signal Processing, **2019, 8, 40-49**
Published Online August 2019 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijigsp.2019.08.04

Visual Object Tracking by Fusion of Audio Imaging in Template Matching Framework

Satbir Singh
Dr.B R Ambedkar National Institute of Technology, Jalandhar, India
Email: satbir2008@ymail.com

Arun Khosla
Dr.B R Ambedkar National Institute of Technology, Jalandhar, India
Email: khoslaak@nitj.ac.in

Rajiv Kapoor
Dr.B R Ambedkar National Institute of Technology, Jalandhar, India
Email: rajivkapoor.dtu@gmail.com

Received: 04 April 2019; Accepted: 24 June 2019; Published: 08 August 2019

***Abstract*—**Audio imaging can play a fundamental role in computer vision, in particular in automated surveillance, boosting the accuracy of current systems based on standard optical cameras. We present here a method for object tracking application that fuses visual image with an audio image in the template-matching framework. Firstly, an improved template matching based tracking is presented that takes care of the chaotic movements in the template-matching algorithm. Then a fusion scheme is presented that makes use of deviations in the correlation scores pattern obtained across the individual frame in each imaging domain. The method is compared with various state of art trackers that perform track estimation using only visible imagery. Results highlight a significant improvement in the object tracking by the assistance of audio imaging using the proposed method under severe challenging vision conditions such as occlusions, object shape deformations, the presence of clutters and camouflage, etc.

***Index Terms*—**Audio Imaging, Template Matching, Object tracking, Information fusion.

I. INTRODUCTION

The video-alone tracking has many constraints especially under challenging visible conditions, such as illumination variation, occlusion, object appearance change, etc. To overcome these limitations, the research trends are moving towards the amalgamation of different sensing domain to aid the visible imagery tracking. New developments for fusion based tracking encourage the use of information through multiple cameras, thermal cameras, audio devices, radio waves, and light waves, etc. The detailed overview can be found in [1, 2, and 3].

We Humans observe a tremendous number of combined visual-audio examples and learn the correlation between them throughout their whole life [4] unconsciously. Because of the correlation between the sound and the visual events, humans are able to understand the object/event that is a sound generator and hence can localize the sound source even without any distinctive education. The motivation for the combined role of audio-visual information by the human in everyday life and its use is provided by the findings in psychology and cognitive science [4, 5, 6, 7, 8, and 9] on sound source localization capability of humans. The findings reveal that visual information correlated to sound improves the efficiency of search [8] and the accuracy of localization [7].

Some major advantages of audio information include Assistance in case of the limited field of view of a camera, providing additional hints in case of occlusion or similar structure clutter presence, Invariant to lighting variations in the scene, robustness, and reliability toward adverse weather circumstances like fog, dust, and snow. Acoustic information is also robust to camouflage for the reason that a sound source can be localized even if it is hidden behind other objects e.g. gunfire in a crowded scene.

The organization of the rest of the paper is presented in this paragraph. The implemented work procedure is explained in detail in Section II. It describes the improved correlation score based template matching used for object tracking in the individual sensing domain followed by the final fusion strategy applied. Section III provides the detail on experimental settings along with the discussion on the results acquired. Section IV highlights the concluding remarks and the future possibilities of the performed research.

II. LITERATURE REVIEW

The purpose of this section is to contemplate upon the previous work done for providing solution to the taken research problem. The study of different state of art techniques helps to understand the problem and its solution in a much organized manner.

Further as the need and significance of the audio-visual is explained earlier in the paper, it becomes important to cover the various practices involved in performing the said bi-modal fusion for the application of object tracking.

Previously, the audio information was found and used in its simple one-dimensional form, in which, the audio information was captured using a number of microphones. Then, certain audio features were obtained from the received sound signal. Finally, the information was used together in conjunction to take the final decision. Several issues like multi-target tracking [10, 11], robustness to reverberation [12], ground noise and interfering sources [13] and time-varying and intermittent sources [14] have been researched, making audio tracking a developed research field.

However, most of the system used for audio tracking are based on a limited number of microphones (typically < 20) spread in the environment. Such kind of layout does not ensure an efficient accuracy of localization and limits the number of sources that can be jointly localized. Further, the relation between the true location of the target and measurements from the microphones is characterized in a non-linear way that forces to use ad hoc tracking algorithms with complex models of measurement noise [15]. In addition, in order to make this information useful for amalgamation with visible mode information, a tedious task of calibration is required to be performed.

More recently, new techniques have been developed to record and generate acoustic images. Microphone arrays are capable of providing spatial information for incoming acoustic waves, as they can capture key information that would be impossible to acquire with single microphones [23]. Acoustic imaging microphone arrays (sometimes referred to as "acoustic cameras" [24]) often contain a camera which is usually located at the center of the array. An acoustic map, spawned using the microphone data, is put over as a transparency over the camera image. The output from acoustic beamforming applied to the signals acquired by a set of microphones encode at each pixel the sound intensity coming from each spatial direction. The devices generating audio imaging in such manner has numerous returns for the video based surveillance, such as the bimodal information can be easily fused together thanks to the pixel-to-pixel correspondence, which in turn diminishes the need of tedious calibration processes. Tracking can be performed taking as a measurement the cues of an image instead of cues of the single microphone's raw signal. The relation between measurement and state in tracking is straightforward. The overlap between acoustic and optical image allows a fair comparison between audio and video trackers. Apart from this, this technology is a completely passive technology, differently from active radars, optical and infrared cameras that require light emitters [25].

The proposed method presents an object tracking method developed specially for an advanced system that eliminates the need for calibration such as presented in [25]. Here, the image-audio arrangement is automatically mapped to an image-image scenario. So, common features can be extracted from both domains. This, in turn, can reduce the complexity of the fusion-based model.

In the last list of various audio-visual fusion based tracking schemes is presented in table 1. This tabulated analysis presents the citation of work, number of sensors used in the process and the comparative relation with the proposed technique that highlights the similarities and the differences among these.

Table 1. Various Audio-Visual Fusion based Tracking Methods

Research	**Sensors used**	**Comparison with proposed research technique**
Volkan Kilic et al. [16]	2 Circular Microphone arrays and 3 cameras	Aid used for object tracking is in the form of direction of arrival of sound and histogram probability matching as compared to the template match score in the proposed method. In addition, projection based homographic calibrations used.
Checka, Wilson, Siracusa, & Darrell [17]	2 cameras and 4 microphone arrays	Their model utilized particle filter framework, in which the foreground and background information was taken as visible cue, but short time Fourier transform of the received audio signal, was taken as the audio cue. So, both the cues were different and then were combined by converting them to probability distribution form.
Gehrig, Nickel, Ekenel, Klee, & McDonough [18]	4 cameras and 16 Microphone in array	This method was intended for an indoor system of a lecture hall with very less disturbance. It used Particle filter for object tracking and 3D projections for obtaining equivalent inputs from both audio and video signals.
Lim & Choi [19]	1 stereo camera and 3 microphones	It used particle filter for forming association between two modes, whereas the proposed method uses correlation based weighted fusion for the combined track estimate.
Megherbi, Ambellouis, Colôt, & Cabestaing [20]	1 camera and 1 microphone	The audio features used were acoustic vectors composed of linear predictive coding coefficients (LPCC) of their speech. In addition, the visible features were based on histogram probability based distributions. Hence, both modality features were diverse which are same in the proposed research. Further, for fusion, Dempster's rule of combination in belief theory was used.
D'Arca, Hughes, Robertson, & Hopgood [21]	1 camera and 16 microphones	Main emphasis was on the condition of occlusion only. It used Kalman filter for fusion, but that is good for linear motion only. It used Time Difference of Arrival feature for audio source localization and height was determined by the visible signal.
Shivappa et al. [22]	4 cameras and 24 microphones	This method used Time Difference of Arrival from different microphone signals and used multi iterative decoding for fusion with visible tracker signal. The major differences were the use of a tough calibration technique that limited this algorithm in terms of speed.

III. Proposed Work Methodology

Initially, the object tracking procedure for an individual domain was performed using a modified robust template matching approach. The simple template-matching algorithm is a popular method for object tracking. However, refinement of the algorithm is required to eliminate the issue of chaotic movements occurring in the template-matching framework. After, the application of amended template matching, the results obtained from individual sensing domain are intelligently fused to obtain the final track estimate. Fig. 1 presents the overall view of the work methodology. The complete section is divided into the following subsections:

A. Tracking using Correlation based Template matching

For tracking in an individual domain, a fast template matching process was incorporated. In this process, the object's initial template was made to slide across the image at the equivalently sized patch interval to find the resultant correlation coefficient values among the different regions of the current frame image. The following mathematical equation provides the formulation for this calculation.

$$s(x,y) = \frac{\sum_i \sum_j (I(x+i,y+j)-\bar{I})(R(i,j)-\bar{R})}{\sqrt{\{\sum_i \sum_j (I(x+i,y+j)-\bar{I})^2\}\{\sum_i \sum_j (R(i,j)-\bar{R})^2\}}} \quad (1)$$

In the above equation, I stands for the current frame image and R for the template region. Here, (i,j) and (x,y) resembles the pixel position in the template region of the object and the image frame respectively. Values, $\bar{R}$ and $\bar{I}$ represent the mean value of pixels in the template region and the corresponding image region with the size equivalent to the template size. The output $s(x,y)$ represent the score or the correlation based template-matching result of a rectangular region in the image with (x,y) as the starting pixel location.

For visible domain tracking, separate measures of R, G, and B were calculated and the combined correlation measure was calculated as the mean of these.

$$s^{VI}(x,y) = (s_r^{VI}(x,y) + s_g^{VI}(x,y) + s_b^{VI}(x,y))/3 \quad (2)$$

In the above equation, $s_r^{VI}(x,y)$, $s_g^{VI}(x,y)$, $s_b^{VI}(x,y)$ represent the individual color channel based correlation similarity score and $s^{VI}(x,y)$ denotes the combined value for visible imagery based output.

The computation from the audio-based sensor is very simple and easier because the information is sufficient by monochrome color representation only. Moreover, only the speaking person's voice gets a corresponding correlation coefficients were found using equation (1) and were resembled as: $s^{AI}(x,y)$.

B. Chaotic Movement Regulation for Template Match Tracking

To avoid superfluous anticipation activity of template matching algorithm, an extra measure was introduced to recheck the obtained track estimate. The state transition in the next frame was certified after verification using a Chaotic Movement Regulation Factor (CMRF). The value of this factor was computed using equation (3). In addition, the check step introduced for eliminating in-consistent tracking is presented by equation (4):

$$CMRF_t = \sqrt{(X_t - X_{t-1})^2 + (Y_t - Y_{t-1})^2} \quad (3)$$

Here, (X_t, Y_t) stands for initial coordinates of rectangular track estimate from the individual tracker at image frame at time interval t. In addition, the corresponding control action was taken was:

$$(X_t, Y_t) = \begin{cases} (X_{t-1}, Y_{t-1}), & if\ CMRF_t > D \\ (X_t, Y_t), & otherwise \end{cases} \quad (4)$$

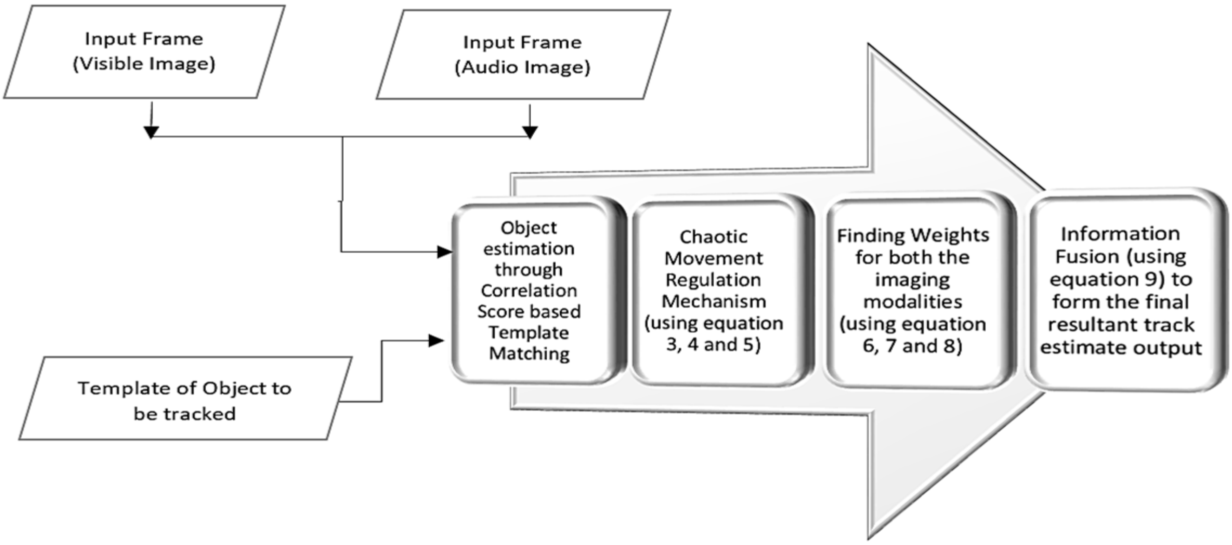

Input Frame (Visible Image)
Input Frame (Audio Image)
Object estimation through Correlation Score based Template Matching
Chaotic Movement Regulation Mechanism (using equation 3, 4 and 5)
Finding Weights for both the imaging modalities (using equation 6, 7 and 8)
Information Fusion (using equation 9) to form the final resultant track estimate output
Template of Object to be tracked

Fig.1. Block Diagram to illustrate the workflow of the proposed scheme

$$D = \sqrt[2]{(A^2 + B^2)} \tag{5}$$

The value of D was estimated using equation (5). It is interesting to note that the threshold parameter D was not chosen as an arbitrary value; rather it was made directly dependent on the dimensions of the object being tracked. Here, A and B represent the width and height of the target selected initially.

After regulating the correlation-based tracking, the final estimates for individual sensing domain were obtained. The maximum value of the correlation coefficient from the above process and the region corresponding to this was chosen as the track estimate from the individual modality. The maximum value of the correlation score for each domain were found among the different correlation values obtained. These were labeled as S_t^{VI}, S_t^{AI} and the region occupying this was saved as state estimates P_t^{VI}, P_t^{AI}respectively.

IV. Fusion Strategy

After the task of improved individual domain tracking estimate had been completed, the next task was to fuse the inputs from mutually exclusive track estimates. For achieving this, effective weights for each complimenting imagery were computed. The variation of the template matching based correlation score pattern throughout the current image frame was chosen as the basis of weighing these sources. The mode having more chances of a unique detection was provided the proportionally greater weight. The value of standard deviation was calculated among the different correlation scores obtained in a complete image frame of individual modality. It is interesting to note that more the pattern of correlation values is varying more are the chances of correct tracking estimate. The adaptation in weights from the individual imaging modality was performed using equations (6) to equation (8).

$$\sigma_t = \sigma_t^{VI} / (\sigma_t^{VI} + \sigma_t^{AI}) \tag{6}$$

$$W_t^{VI} = \frac{\sigma_t \; S_t^{VI}}{(\sigma_t \; S_t^{VI} + (1-\sigma_t) \; S_t^{AI})} \tag{7}$$

$$W_t^{AI} = \frac{(1-\sigma_t) \; S_t^{AI}}{(\sigma_t \; S_t^{VI} + (1-\sigma_t) \; S_t^{AI})} \tag{8}$$

In the above equations σ_t^{VI} and σ_t^{AI} represent the standard deviation value for the patterns of correlation scores obtained for individual imaging domain at the current time t. The tuning parameter for adjustments of weights was found using equation (6) and is denoted as σ_t. Next, the weights to these sources are provided accordingly using equation (7) and (8). Here, the parameters W_t^{VI} and W_t^{AI} denote the weighting factors for individual sensing imaging.

$$P_t^F = W_t^{VI} P_t^{VI} + W_t^{AI} P_t^{AI} \tag{9}$$

The final tracking estimate P_t^F is obtained using the earlier found tracking output of individual imaging modality values, P_t^{VI} and P_t^{AI}.

V. Experimental Results & Discussion

The testing of the proposed framework for audio imaging assisted video-based object tracking was done on three sequences available at [26]. The details of these sequences are presented in Table 2. The results along with discussion are presented according to the data sequence number. The performance of the Proposed Audio – Visual Fusion algorithm (PAVF) is compared with the alone visible trackers such as Visible alone Template Matching (VTM) and Visible Particle Filter [27] (VPF) algorithm. Along with this, the performance is also compared with a multi-cue vision based method that presents the fusion using intensity and gradient features (VCGF) [28]. All the experiments were performed using MATLAB software. The quantitative analysis is performed using some statistical parameters, whose detail is presented in the following paragraph.

The parameters opted for comparing the efficiency of the proposed algorithm are Root Square Positional Error Value (RSPEV), Precision Value (PV), Recall Measure Value (RMV) and finally the F1-Measure Value (FMV). The comparisons were obtained by taking the object's ground truth positions in the video sequence. The RSPEV resembles the closeness in the position of the track estimate with the actual position of the target, whereas PV and RV are the resultant of bounding box overlap efficiencies. PV is found as a ratio of intersection area of ground-truth bounding and estimated track bounding box to the area covered by the later. Whereas for RV, the numerator in the ratio remains as it is but the denominator is exchanged by the area of the ground truth bounding box. FMV is the combined measure that is equal to the Harmonic mean of PV and RV. F-score can be criticized in particular circumstances due to its bias as an evaluation metric. The mathematical calculations of these parameters are provided in the following four equations:

$$RSPEV_t = \sqrt{\left(G_t(x) - P_t^F(x)\right)^2 + \left(G_t(y) - P_t^F(y)\right)^2} \tag{10}$$

$$PV_t = G_t(area) \cap P_t(area) / P_t(area) \tag{11}$$

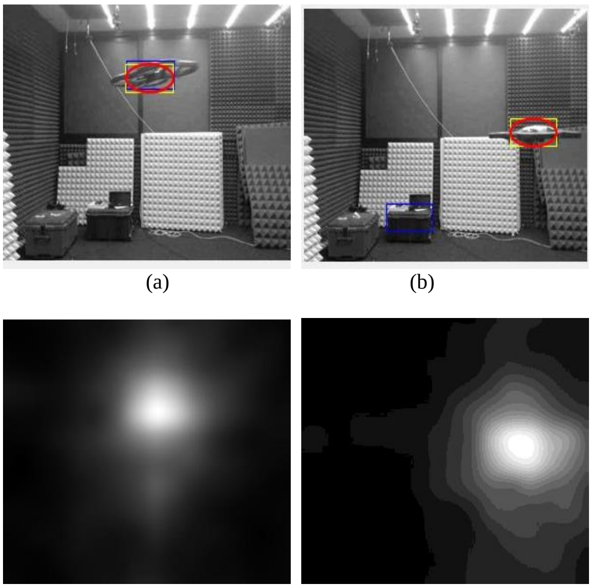
$$GV_t = G_t(area) \cap P_t(area) / G_t(area) \tag{12}$$

$$FMV_t = 2 \times P_t \times R_t / (P_t + R_t) \tag{13}$$

In above equations, $G_t(x)$ and $G_t(y)$ are the X and Y coordinate of starting point of the actual target state respectively, whereas, $P_t^F(x)$ and $P_t^F(y)$ represent the X and Y coordinate of starting point of the final track estimate found through applied fusion algorithm.

A. Sequence1:

The scenario for tracking the drone in the air can be observed in Fig. 2. The two frames illustrate that the drone appearance changes as it moves across the video sequence. In Fig. 2(a), some of the bottom view of the Drone is captured, whereas Fig. 2(b) mainly presents the side view of the Drone. Hence, it becomes difficult for any of the visual feature based object trackers to identify the drone with changing appearance. Moreover, throughout the whole sequence, the drone occurs many rotations, translations as well as its appearance changes in the visible domain.


(a) (b)

(c) (d)

Fig.2. Illustration of Sequence 1: (a), (b) Frame 6 and Frame 100 of Visible Image of the sequence, (c), (d) Audio Imaging counterpart of (a) and (b) respectively.

But, the Audio Imaging produced is unaffected by these changes as shown in Fig. 2 (c) and Fig. 2(d). Therefore, in this case, the audio imaging becomes a more reliable source for tracking as compared to its visible counterpart. Hence, the proposed algorithm adjusts the weight favoring under these challenging conditions. The weight provided to the incorrect estimation of visible template matching tracker (shown in Blue rectangle in Fig. 2(c)) for frame number 100 is 0.095, whereas the weight provided for the same case to the audio image has a value of 0.905. In Fig. 2, the area covered under the yellow rectangle represent the track output from the audio imaging, whereas the red circle encloses the final output track estimate.

The same fact can be also be verified by going through the quantitative parameter analysis. Where the rest of the algorithm fails continuously from starting frame number 6 until the last frame number 150. Here, the analysis was started from frame number 6, because an appropriate template for matching was found in frame number 5 of the video. Hence, the tracking process was started thereafter. The plot of RSPEV for Drone sequence in Fig. 3 indicates that the audio imaging assisted tracking outperforms various visible tracking methods.

B. Sequence 2

This sequence is more difficult than the drone sequence because apart from the rotations and changing appearance of the object, the speaking person to be tracked comes across various partial and full occlusions in the visible domain. Further, there are similar visible appearance objects present in the scene.

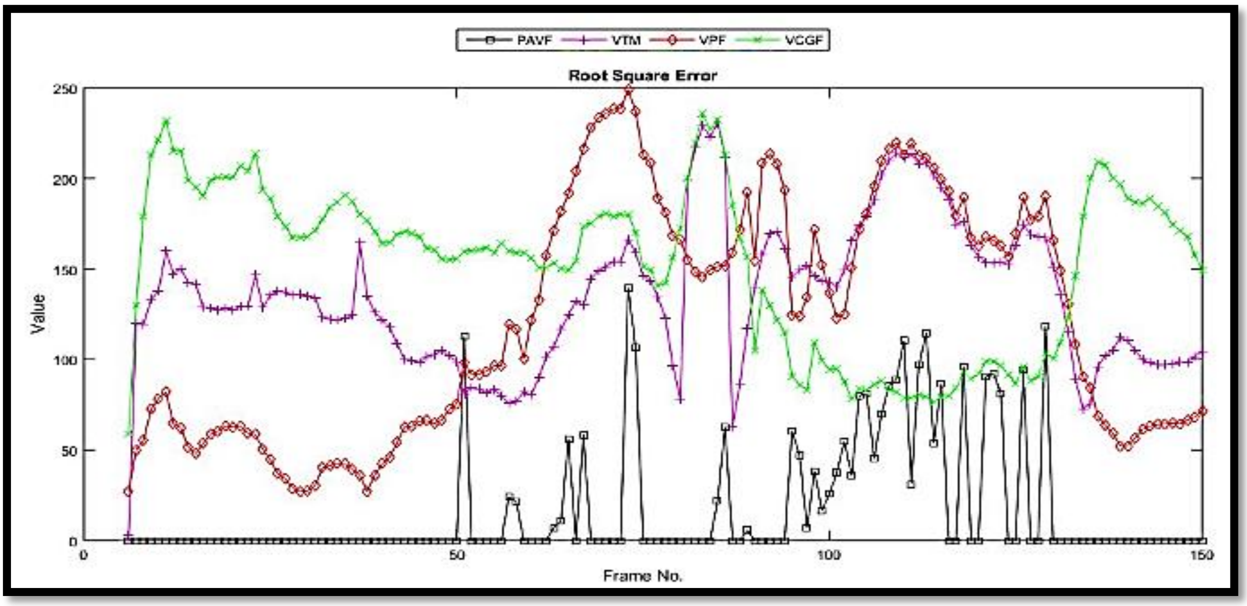


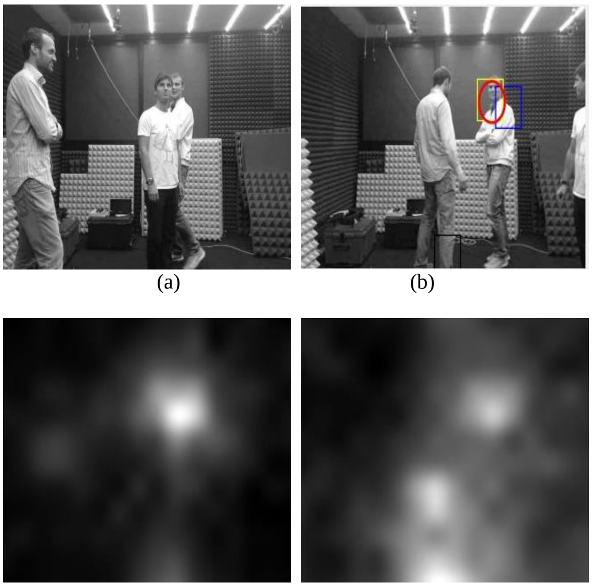
Fig.3. Plots for RSPEV comparisons for Sequence1

Table 2. Type Details of Dataset Sequence taken for experimentation

Dataset	Remarks	Challenges
Sequence 1 [26]	A 2 minutes long sequence, taken in a moderately reverberant environment. The goal is the tracking of a drone flying in the room, whose propellers are the only active audio sources.	The drone itself goes certain deformations in its size and shape during motion.
Sequence 2 [26]	A 2 minutes long sequence, taken in a moderately reverberant environment. The goal is the tracking of the face of a speaker moving in the room.	Presence of disturbing video and audio sources, generated by other people moving in the room
Sequence 3 [26]	Taken outdoor from a terrace, looking at a road about 50 m far from the device. The goal is the tracking of a motorbike.	Many disturbing sources are present, including wind flurries causing trees movements, a highway at about 500 m from the camera in the upper zone of the image

The video sequence is illustrated with the help of Fig. 4. The frame 10 of this visible-acoustic imaging dataset (presented in Fig. 4(a) and Fig. 4(c)) depicts that the partial occlusion condition does not seem to be a hurdle for the audio imaging. The object in the left image can be easily and robustly predicted from the right side acoustic image. Therefore, the fusion algorithm here can make the choice of heavily relying on the clearer differentiable image.

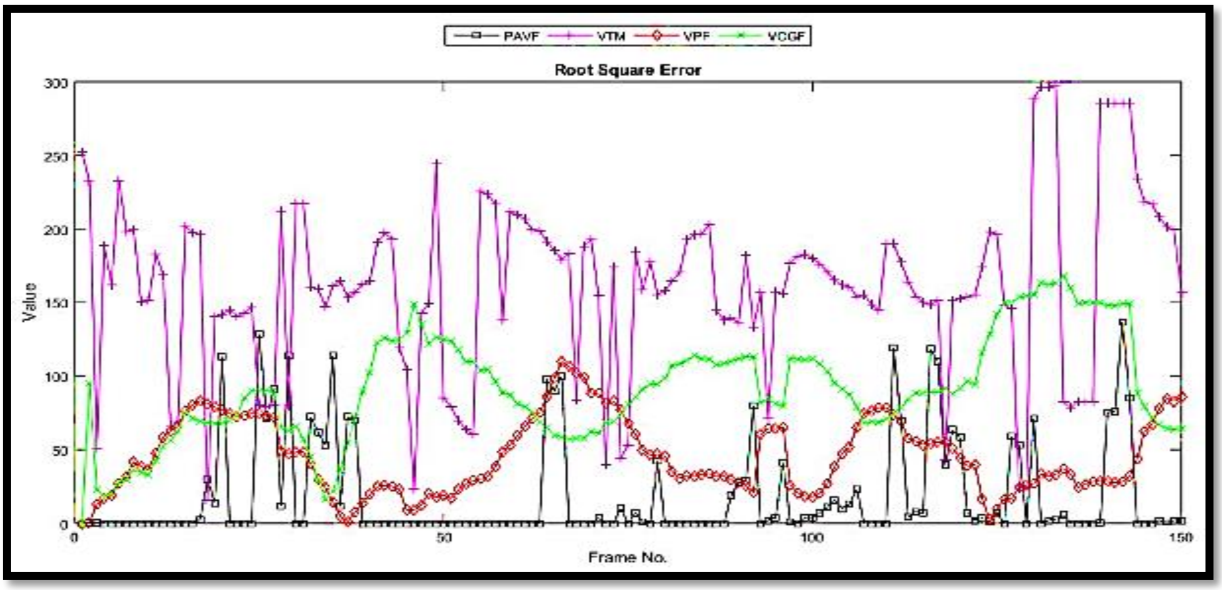
The case of Fig. 4 (b) and Fig. 4(d) present frame number 48 of sequence 2. In this, it is interesting to note that visible template matching is close to the location of the object but not pointing truly to the actual position (blue rectangle). Also, the audio imaging obtained is corrupted by random noise disturbances present. But, due to the mechanism provided in the proposed scheme for regulating the chaotic movements for template match using equation (3-6), the audio imaging tracker overcomes the effect of noises present and still provides a more reliable track estimate shown by an area under yellow rectangle. The weight values provided for this case are 0.36 for the visible tracker and 0.64 for the audio tracker. Finally, the final state estimate is enclosed by the red circle.



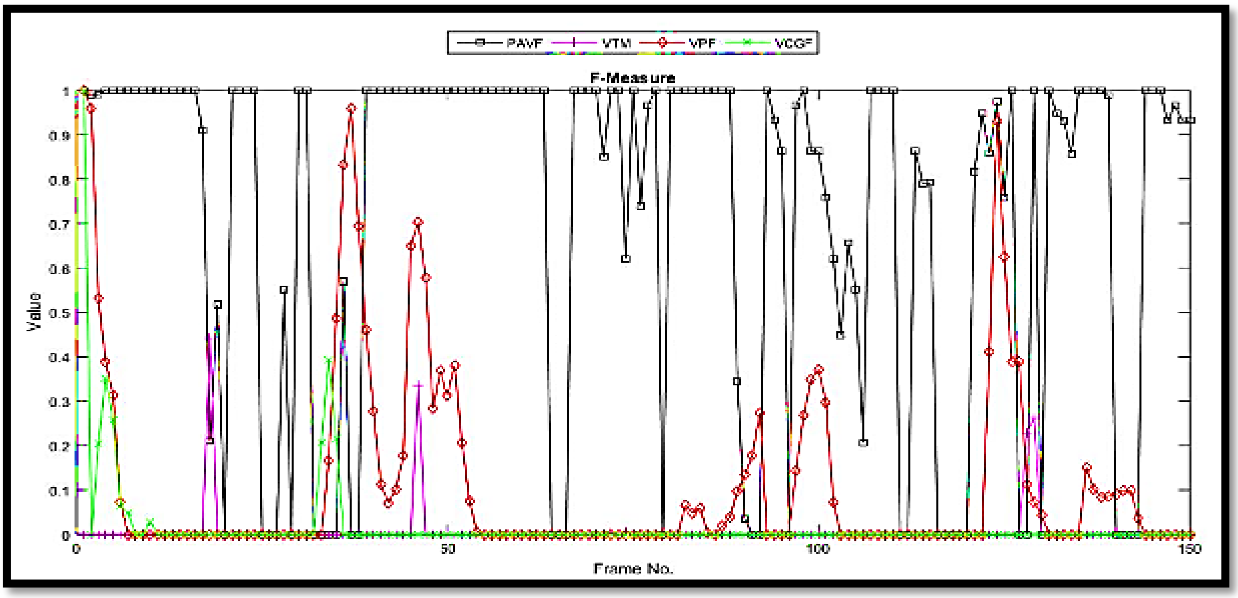
(a) (b)

(c) (d)

Fig.4. Illustration of Sequence 2: (a), (b) Frame 10 and Frame 48 of Visible Image of the sequence, (c), (d) Audio Imaging counterpart of (a) and (b) respectively.



(a)



(b)

Fig.5. Plots for quantitative comparisons for Sequence2: (a) RSPEV and (b) FMV

The performance of the proposed algorithm was compared quantitatively and the resulting plots are presented in Fig. 5. Figure 5(a) illustrates the PRSE values. It is easy to deduce from this plot with the proposed method outperforms the other trackers. The assistance of Audio Imaging played an important role in improving the performance as explained in the above paragraphs. The proposed method also occur little errors in certain time intervals, where both the audio image as well as the video image is not providing cleaner hints in the sequence. Although the chaotic movement is controlled which results in small PRSE values during these difficult conditions, where the rest of algorithms fail miserably. The bounding box overlap ratio was formulated using F1-Measure calculations and Fig. 5(b) presents the obtained plot pattern.

C. *Sequence 3*

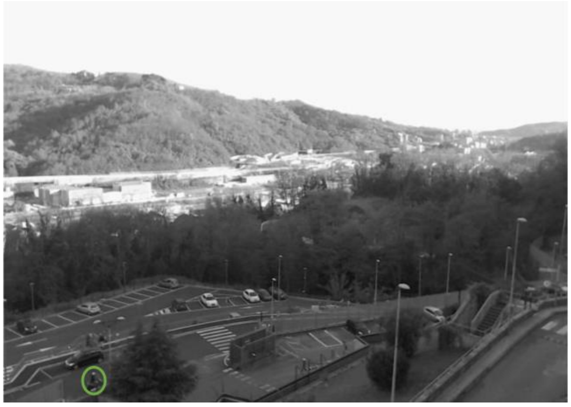
This sequence provides very tough challenges to the tracker. Fist, the size of the object to be tracked is very small as compared to the scene dimensions. Here, the scene is captured from a very far distance. The motorbike to be tracked is circled in green color in Fig. 6(a).

But for this frame, the audio image is comparatively better. It is shown in Fig. 6(c). The scenario in visible imagery also inhibits tough challenges such as a small-sized object, partial and full occlusions, Object rotation, the presence of clutters and similar appearance objects.

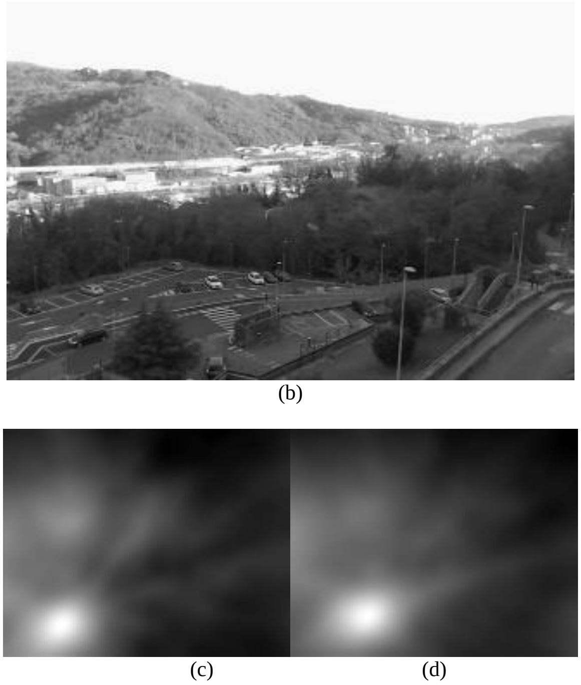
Also in Fig. 6(b), it is observed from the naked eye that the object has disappeared from the scene. But, actually, the biker gets occluded by the tree in its path. But, for acoustic imaging based signal, the presence of the object is relatively clear and is shown by Fig. 6(d). So, the single domain tracking would have been surely a failure for this type of sequence.

Apart from these, the audio signal here also encounters vague noises in bulk due to open highway recording, with sounds from vehicles and people present on the road. Since the recording was done on a terrace (an open noisy environment), it further worsened the situation for proper audio imaging. Hence, for this imagery, the results are not that promising as compared two earlier sequences.

But, one thing was surely determined that the assistance of audio imaging through the proposed fusion based tracking helps to resolve some of the major challenges for the vision-alone based object tracking. Table 3 and Table 4 provide the values of Average value of the statistical parameters per data sequence for providing the comparative analysis among different algorithms.



(a)



(b)

(c) (d)

Fig.6. Illustration of Sequence 3: (a), (b) frame 15 and frame 44 of Visible Image of the sequence, (c), (d) Audio Imaging counterpart of (a) and (b) respectively

Table 3. PRSE values for different algorithms

Algorithm	Sequence1	Sequence2	Sequence3	Average
PAVF	**13.24**	**18.12**	**35.14**	**22.17**
VTM	52.92	115.77	135.97	101.53
VPF	45.90	62.40	82.40	63.57
CGF	61.84	83.53	93.61	79.66

Table 4. F-Measure Score for various Algorithms

Algorithm	Sequence1	Sequence2	Sequence3	Average
PAVF	**0.81**	**0.78**	**0.66**	**0.75**
VTM	0.20	0.13	0.10	0.14
VPF	0.32	0.23	0.19	0.25
CGF	0.22	0.19	0.13	0.18

Here, the top performing algorithm is marked in bold letters. For PRSE, minimum value and for F-measure, maximum value is considered as the most efficient outcome respectively.

VI. Conclusion

We presented here a new development for object tracking using the audio imaging – visual imaging fusion in the template matching framework. Audio imaging can play an important role in computer vision, particularly for automated surveillance by boosting the accuracy of present systems. It has a great advantage in comparison to the primitive audio signal mode of information because it eliminates the need for tedious calibration process. In this paper, an improved template matching based tracking was presented that took care for the chaotic movement of the ordinary template-matching algorithm. Along with it, a fusion scheme was proposed that utilized the standard deviation values obtained from the correlation score pattern obtained throughout the individual frame in each imaging modality. Upon comparison with the state of art visible trackers, it was evident from the results that the proposed method brought considerable improvement in the object tracking process. The accuracy was judged in terms of positional errors and overlap efficiency. The lowest positional error with average value of 22.17 (compared to 101.53, 63.57, and 79.66) and highest overlap measure of 0.75 (compared to 0.14, 0.25, and 0.18) justified the contributions made by our method. Moreover, the performance of the algorithm in different visible conditions such as deformations in its size and shape (sequence 1), presence of disturbing video and audio sources, generated by other people moving in the room (sequence 2), and a difficult visible scene along with many disturbances (sequence 3) was comparatively appreciable. Although, the proposed method outperformed state of art methods in all the sequences, but there is still a scope for improvement, especially for the cases of presence of hefty noises (as was in the case of the third sequence of the outdoor motorbike). However, one can definitely say that conversion of sound to the imaging domain automatically is a plus point and for further work, various image enhancements techniques like image equalization algorithms and image filtering techniques (spatial domain, frequency domain, wavelet domain etc.) can be applied in the future for bringing further robustness to the process of Object Tracking.

References

[1] G. S. Walia and R. Kapoor, “Recent advances on multicue object tracking: a survey,” *Artificial Intelligence Review*, vol. 46, no. 1, pp. 821–847, 2016.

[2] S. Singh, R. Kapoor, and A. Khosla, *Cross-Domain Usage in Real Time Video-Based Tracking*. U.S.A: IGI Global, 2017, pp. 105–129.

[3] J. Ma, Y. Ma, and C. Li, “Infrared and visible image fusion methods and applications: A survey,” *Information Fusion*, vol. 45, pp. 153–178, 2018.

[4] W. W. Gaver, “What in the world do we hear: An ecological approach to auditory event perception,” *Ecological psychology*, vol. 5, no. 1, pp. 1-29, 1993.

[5] B. Jones and B. Kabanoff, “Eye movements in auditory space perception,” *Perception, & Psychophysics*, vol. 17, no. 3, pp. 241-245, 1975.

[6] P. Majdak, M. J. Goupell, and B. Laback, “3-d localization of virtual sound sources: Effects of visual environment, pointing method, and training,” *Attention, Perception, & Psychophysics*, vol. 72, no. 2, pp. 454-469, 2010.

[7] B. R. Shelton and C. L. Searle, “The influence of vision on the absolute identification of sound-source position,” *Perception & Psychophysics*, vol. 28, no. 6, pp. 589-596 1980.

[8] R. S. Bolia, W. R. D’Angelo, and R. L. McKinley, “Aurally aided visual search in three-dimensional space,” *Human Factors*, vol. 41, no.4, pp. 664-669, 1999.

[9] D. R. Perrott, J. Cisneros, R. L. McKinley, and W. R. D’Angelo, “Aurally aided visual search under virtual and free-field listening conditions.” *Human Factors*, vol. 38, no.4, pp. 702-715, 1996.

[10] M. F. Fallon and S. Godsill, “Acoustic source localization and tracking using track before detect,” *IEEE*

Transactions on Audio, Speech, and Language Processing, vol. 18, no. 6 pp. 1228–1242, 2010.

[11] A. Plinge and G. Fink, "Multi-speaker tracking using multiple distributed microphone arrays," in *IEEE Int. Conf. on Acoust., Speech and Sig. Proc.* , May 2014, pp. 614–618.

[12] D. B. Ward, E. Lehmann, and R. Williamson, "Particle filtering algorithms for tracking an acoustic source in a reverberant environment," *IEEE Trans. On Speech and Audio Proc.*, vol. 11, no.6, pp. 826–836, 2003.

[13] K. Wu, S. T. Goh, and A. W. Khong, "Speaker localization and tracking in the presence of sound interference by exploiting speech harmonicity," *in IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 365–369.

[14] E. A. Lehmann and A. M. Johansson, "Particle filter with integrated voice activity detection for acoustic source tracking," *EURASIP Journal on Applied Signal Processing*, vol. 2007, no. 1, pp. 28–28, 2007.

[15] V. Cevher, R. Velmurugan, and J. H. McClellan, " Acoustic multi-target tracking using direction-of-arrival batches," *IEEE Transactions on Signal Processing*, vol. 55, no. 6, pp. 2810–2825, 2007.

[16] V. Kilic, M. Barnard, W. Wang, and J. Kittler, "Audio Assisted Robust Visual Tracking With Adaptive Particle Filtering," *IEEE Transactions on Multimedia*, vol. 17, no. 2, pp. 186–200, 2015.

[17] N. Checka, K. W. Wilson, M. R. Siracusa, and T.Darrell, "Multiple person and speaker activity tracking with a particle filter," *IEEE International Conference on Acoustics Speech and Signal Processing*, (1), V-881-4. 2004.

[18] K. Nickel, T. Gehrig, R. Stiefelhagen, and J. McDonough, "A joint particle filter for audio-visual speaker tracking," in proceedings of the 7th International Conference on Multimodal Interfaces, 2005.

[19] Y. Lim and J. Choi, "Speaker selection and tracking in a cluttered environment with audio and visual information," *IEEE Transactions on Consumer Electronics*, vol. 55, no. 3, pp. 1581–1589, 2009.

[20] N. Megherbi, S. Ambellouis, O. Col ôt, and F. Cabestaing, "Joint audio-video people tracking using belief theory," in *IEEE International Conference on Advanced Video and Signal Based Surveillance*, p. 135–140, 2005

[21] E. D'Arca, A. Hughes, N. M. Robertson, and J. Hopgood, "Video tracking through occlusions by fast audio source localization," in *IEEE International Conference on Image Processing*, pp. 2660–2664, 2013.

[22] S. T. Shivappa, B. D. Rao, and M. M. Trivedi, "Audio-visual fusion and tracking with multilevel iterative decoding: Framework and experimental evaluation," *IEEE Journal on Selected Topics in Signal Processing*, vol. 4, no. 5, pp. 882–894, 2010.

[23] F. Su, "Acoustic Imaging Using a 64- Node Microphone Array and Beamformer System," thesis submitted to Carleton University Ottawa, Ontario, 2015.

[24] M. Legg, and S.Bradley, "A combined microphone and camera calibration technique with application to acoustic imaging," *IEEE Transactions on Image Processing*, vol. 22, no. 10, pp. 4028-4039, 2013.

[25] A. Zunino et al., "Seeing the Sound: a New Multimodal Imaging Device for Computer Vision," in *IEEE conference on Computer Vision Workshop*, pp. 693-701, 2015.

[26] http://www.iit.it/en/pavis/datasets/DualCam.html

[27] K. Nummiaro, E. K. Meier, and L. V. Gool, "An adaptive color-based particle filter," *Image Vis. Comput.*, vol. 21, no. 1, pp. 99–110, 2003.

[28] J. Xiao, R. Stolkin, M. Oussalah, and A. Leonardis, "Continuously Adaptive Data Fusion and Model Relearning for Particle Filter Tracking With Multiple Features," *IEEE Sens. J.*, vol. 16, no. 8, pp. 2639–2649, 2016.

Authors' Profiles

Satbir Singh Satbir Singh received M.E in Electronics and Communication Engineering from Thapar University, Patiala, India. He is presently working with Central scientific Instruments Organization, Chandigarh, India. Previously, he has a working experience as Senior Scientific Officer with Electronics and Communication Engineering Department of Delhi Technological University, Delhi and as a Project Engineer with Centre of Advanced Computing, Mohali, India. His research interests include signal processing, computer vision and IoT. Currently, he is pursuing Ph.D. from National Institute of Technology, Jalandhar, India.

Arun Khosla Arun Khosla received his Ph.D. degree from Indraprastha University, Delhi in the field of Information Technology. He is presently working as Professor in the Department of Electronics and Communication Engineering, National Institute of Technology, Jalandhar, India. Dr. Khosla has been reviewer for various IEEE and other National and International conferences and also serves on the editorial board of International Journal of Swarm Intelligence Research. He is a life member of Indian Society of Technical Education.

Rajiv Kapoor Rajiv Kapoor received M.E. and Ph.D. degree in Electronics & Communication Engineering from Delhi College of Engineering, Delhi University. Dr. Kapoor presently working as Professor in Electronics & Communication Engineering Department, AIACT&R (Govt. of NCT of Delhi). He has authored over 90 research papers in various renowned international journal and conferences. His primary research interests are machine learning, computer vision, signal and image processing.

How to cite this paper: Satbir Singh, Arun Khosla, Rajiv Kapoor, " Visual Object Tracking by Fusion of Audio Imaging in Template Matching Framework", International Journal of Image, Graphics and Signal Processing(IJIGSP), Vol.11, No.8, pp. 40-49, 2019.DOI: 10.5815/ijigsp.2019.08.04