I.J. Image, Graphics and Signal Processing, **2023, 4, 1-18**
Published Online on August 8, 2023 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijigsp.2023.04.01

Object tracking via a Novel Parametric Decisions based RGB-Thermal Fusion

Satbir Singh*
Dr B R Ambedkar National Institute of Technology/ Centre for Artificial Intelligence, Jalandhar, 144088, India
E-mail: satbirs.ai@nitj.ac.in
ORCID iD: https://orcid.org/0000-0002-6263-2092
*Corresponding Author

Arun Khosla
Dr B R Ambedkar National Institute of Technology/ Centre for Artificial Intelligence, Jalandhar, 144088, India
ORCID iD: https://orcid.org/0000-0001-8571-7614

Rajiv Kapoor
Delhi Technological University, Delhi, 110042, India
ORCID iD: https://orcid.org/0000-0003-3020-1455

Received: 08 September, 2022; Revised: 02 November, 2022; Accepted: 25 January, 2023; Published: 08 August, 2023

Abstract: The thermo- visual fusion based tracking has been deployed for overcoming the shortcomings of alone vision-based object tracking. The assistance from both domains should be wisely merged so that it should result in a useful practice for object tracking. Several techniques had been developed recently to implement a brilliant fusion, but this undeveloped field still inhibits many unsolved challenges. The proposed method aims at increasing the effectiveness of tracking by bi-modal fusion with the introduction of a new set of rules based upon the parameters generated from the decision of individual modality trackers. This practice helps to achieve output by only a single run of the fusion process in every frame. The method also proposes to use minimal information from individual trackers in normal conditions and incorporates the use of supplementary information from imageries merely in case of diverse working conditions. This procedure, in turn, lessens the computations and hence reduces time to process. The experiments performed on well-known publically available datasets show the advantages of the proposed method over the individual visual domain tracking and other existing states of the art fusion techniques.

Index Terms: Particle filter, Object tracking, Decision level fusion, visible-thermal amalgamation.

1. Introduction

Object tracking using sensor fusion of thermal data and the coloured video is a more efficient approach as compared to tracking in isolated individual domains. Visible camera mainly depends on the colour content of the object and surroundings. Thermal / IR camera only captures the thermal image of the object. The thermal image is exclusively related to the temperature value of the object. High-temperature objects possess precious intensity values, whereas low-temperature objects do not possess significant intensity in a thermal image. This information is beneficial for surveillance applications. Moreover, in the case of human target tracking, background illumination variations and inconsistent cloth texture can act as difficulties in vision-based monitoring, but the thermal domain imagery is unaffected from it.

On the other hand, the visible camera has its advantage by having more variations in the data acquired through the colour content information. It is also helpful when a similar temperature object comes in contact with one another, and it is difficult to differentiate their thermal distributions. Hence, during disguised conditions and under the presence of challenging complications, both can complement each other effectively. Since both the domains lie in the image plane, it is easiest to calibrate the information obtained from these cross domains. The following subsection introduces the readers with the previous work done in this motivational direction.

1.1. Related Work

Various research studies have been exercised in this sensor fusion technology. Face recognition inhibiting thermal and visible sensor fusion using background subtraction had been proposed earlier in literature [1-4]. It has also been actively utilized in robotics based vision applications [5-7]. Besides this, the use of visible and thermal imagery in fusion has also been hosted in spacecraft proximity operations [8] to overcome the effect of different lightening conditions in the orbital environment. [9,10] have advanced the field of image fusion of multisensory data for the improved computer vision applications. More can be explored in the study of [11,12].

Notably, the use of said bi-modal information has been increased for object tracking under challenging conditions of vision in recent times. Nowadays researchers have been trying to advance this overwhelming field by the application of intense methods such as convolutional neural networks [13] that can automatically extract the deep features. A two-stream convolution net was formed independently for each domain and the fusion was achieved by developing a fusion net. Despite an exciting approach, it is to be noted that it is dependent on a rigorous training process through the use of image net database. For obtaining the instantaneous tracks, the process cannot track unless it has been prepared for feature extraction through a complicated process.

Extending a primary particle filter [14] method for adaptive fusion at particle level [15] was a good idea for tracking under similar background conditions. It introduced an additional parameter in fusion to avoid the camouflaging effect. However, the method lacked in several key aspects such as having a long processing time and repetition of fusion for every particle per each frame. Also, if the total number of particles is increased initially, then computational efficiency is affected and time to process rises severely.

A basic approach to tracking by fusion [16] involved multiplication of the individual specialists that used cues from thermal as well as visible imagery to form their opinions. However, simple multiplication only serves the purpose of video sequences with no veiling conditions. In case of false tracking in one mode of imagery, it may lead to overall false result due to the absence of procedure to verify this.

[17] presented a method in which the fusion was performed single time per frame due to the execution of some weighted fusion rules. In this method, individual tracking was performed in two different algorithms. Particle filter was incorporated for vision/colour and correlation based template matching was applied for thermal counterpart. However, this method also lacked in two significant aspects. First, any one domain is over weighted (>1) and other is under weighted (<1) in case of fusion rule 2 and fusion rule 3 in the paper, which are called for most of the time. The over dependency on one domain is based upon the only difference in maximum likelihood obtained from individual trackers dealing in foreground tracking only, which may be either false or true under challenging conditions. Second, this decision based strategy does not provide any correction in output due to the position of the least weighted imagery domain output, which at long distance may deviate itself and thus affects the location of final estimate through the presented fusion.

For the knowledge of readers, different methods that incorporated thermo-visible fusion based tracking are listed under Table1. The proposed work inspires to resolve the above-said issues of fusion, and the detailed contributions and motivation for the research are mentioned in the following subsections.

1.2. Motivation for the Work

A significant motivation for the research work lays in working out a technique that may be able to make a robust track outcome estimate by fusion of the decision made by individual trackers. So, the repetition of the same fusion rule per single step of tracking procedure should be avoided. The fact is that various existing approaches in literature employ the fusion either at the pixel level of the image or at the feature level of individual particles present in the particle filter. Therefore, both approaches in the categories mentioned above seem to be inefficient at sinking the computing time because the same fusion strategy is executed several times for finding out a single frame track outcome. If it is a pixel level fusion, then the fusion process repetition rate becomes equal to the size of the image, i.e., no. of pixels in the image.

Moreover, in the case of the most commonly used particle filter based tracking, fusion is applied for P turns per single tracking outcome. Here P denotes the total no. of particles used for tracking. So, computation cost related to the fusion portion of the tracking algorithm increases with an increase in the number of particles used. Moreover, some methods in literature [14,18] involve the role of background both in the thermal and visible imagery at almost every frame. This calculation of background region, its feature extraction and then the estimation of fusion parameters are repeated for each particle per single frame analysis, which makes it an ample time consuming task. So, the purpose of this research is to formulate a method that can amalgamate the role of both modalities with temporal efficiency and still provide an accurate track estimate. Another inspiration comes from the fact that significant work is done given likelihood adaptation of individual imagery, but work needs to be done in an adaptation of the role of the single state position estimate in a decision-based fusion framework.

1.3. Contribution of work

The proposed method stays with the use of single feature based particle filter tracking in individual domains but introduces an intelligent fusion framework that can overcome the limitations of either of the individual imagery trackers and results into a credible track estimate. The critical contributions of the proposed research work are mentioned in the following points:

i. The proposed work introduces an efficient design framework to obtain track estimates by using a fusion of thermal and visible imageries. This presented fusion strategy makes it possible that the execution of the fusion step for imageries is not done more than once per single frame as compared to many times per frame execution in previous algorithms. This type of execution, in turn, saves much computational processing.

ii. Second, it presents a selective dealing perspective framework for the tracker. The tracker can either chose to perform simple fusion procedure under normal conditions or can decide to take the supplement help from imageries for making fusion decision only in the presence of diverse challenging circumstances.

iii. A formula is worked out for adjustment of positional coordinates of fused track output after decisive fusion process depending on the outcome provided by the output parameters of individual trackers. This helps in remedying the side effect of hazards in state estimate due to less voted tracker.

iv. A new occlusion treatment is also proposed that is based upon the introduced fusion process. Here, the occlusion detection is the result of the merged opinion from both the imageries.

The organization of the paper is as: Section 2 provides a description of work methodology adopted containing the details of the individual domain particle filter tracking followed by the detailed explanation of the designed fusion strategy along with all rules and parts. The latter part of this section mentions the procedure used while handling occlusions. The following section provides the experimental part along with the discussion on the results acquired covering the qualitative and quantitative aspects. Section 4 contains the concluding remarks and the future possibilities of the performed research.

2. Work Methodology Description

The following section describes the complete details of the work procedure followed for this technique of tracking. The whole process is divided into two main steps. The first subsection provides the insight of the particle filter algorithm incorporated for individual domain tracking. The following subsection provides details of the implementation of the proposed fusion algorithm to obtain the final track estimate.

2.1. Particle Filter Tracking

The implemented particle filter algorithm can be understood appropriately by breaking down the procedure into five stages. Each stage (except initialization of particle filter) was repeated for every frame of video. The explanations of these are as following:

2.1.1. Initializing Particle filter

Initially, on the availability of the target template, its features for reference were generated and fed into the system to act as a reference during the tracking process. To begin with the working of the particle filter, particles of the particle filter were distributed randomly across the scene frame around the target. Each particle was described using a state vector corresponding to $[X, Y, A, B, V_x, V_y]$. In this, X, Y correspond to starting coordinates of the rectangle shaped particle. A, B denote the width and height of the rectangle respectively. V_x and V_y are the velocities, with which, the particles move through the sequence in X and Y direction respectively.

2.1.2. State updating using motion model

After the start, circulated particles states were modified in every frame using a state updating procedure. A linear update matrix, U was used to make the increment in each proceeding frame. Equation (1) describes the state update process for the particles:

$$s_t^p = U \times s_{t-1}^p + G_t \quad (1)$$

Here, s_t^p stands for the p^{th} particle state at time instant $t = t$ and U is taken as (6X6) Identity Matrix, and $G_t \overset{\text{def}}{=} N(0, Q)$. Q is a zero mean Gaussian process with varianceσ_g, whose value is chosen as 0.25 during the experiments.

2.1.3. Observational features from particles

After creation and then updating particles, the next step was to extract information from the state of particles in every frame. Since the fusion algorithm uses two significant aspects: one is weight value and another the positional coordinates of the tracking outcome; therefore, two main features were extracted from every particle. The weight of the particle was awarded based upon a similarity measure calculation.

Table 1. Previous work related to combined Thermo-Visual object tracking

Reference	Algorithms used for tracking	Fusion Strategy	Application
C.Li [13]	Deep Convolutional Networks	Fusion net	RGB-Thermal Tracking
M. Talha et al. [14]	Particle Filter	Adaptive Foreground-Background relation based fusion	Tracking in camouflaging conditions
O'Conaire et al. [15]	Mean Shift tracker	Product of Expert Beliefs	Object tracking
G. Xiao et al. [16]	Particle filter, template matching	Weighted Fusion Rules	Object tracking
Stolkin et al. [19]	Particle Filter	Bayesian Fusion	Human Detection and Tracking
E. Fendri [20]	Background modelling	Logical AND between the binary mask regions in separate imageries	Moving object detection
S. Singh [21]	Particle Filter	Granular computing based fusion	Object tracking in datasets with difficult vision conditions
Y. Niu [22]	Frame difference based background subtraction	Fusion based on target region segmentation	Unmanned Aerial Vehicles (UAV)
S. Singh [23]	Template matching	Correlation based weighted fusion	Object tracking
C. Li [24]	Bayesian Filtering	Adaptive fusion using Collaborative Sparse representation	Tracking in Grayscale-Thermal framework
C. Li [25]	Bayesian Filtering	Laplacian sparse representation	Tracking in Grayscale-Thermal framework
H. P. Liu [26]	Particle Filter Tracking	Using Min operation on Sparse representation	OCTVBS dataset
Palmerini et al. [8]	Kalman Filter	Extended Kalman Filter	Spacecraft Proximity Tracking under light varying conditions

The use of Bhattacharya similarity measure was incorporated to associate the colour distributions of the template of the target to be tracked and the regions estimated from the state of different particles. First, colour histogram with eight bins for each R, G, and B components was found for both the template chosen as well as the region of the image associated with the individual particle in the visible domain. However, the thermal domain made use of only 1D grey intensity histogram. If b denotes the number of bins in the histogram, then similarity measure BS^p for particle p was given by (2) as:

$$BS^p = \sum_{k=1}^{b} \sqrt{T(k) \times R(k)} \tag{2}$$

Here T denotes the template RGB histogram and R stands for the particle region histogram. However, the distances were transformed to fit in the non-linear probabilistic robustness using (3) to weigh the individual particles.

$$w^p = \frac{e^{\{(BS^p)^2 /2\sigma^2\}}}{\sqrt{2\pi\sigma}} \tag{3}$$

In the above equation, w^p stands for weight of p^{th} particle and σ is the variance parameter, whose value is taken as 0.01 throughout the experimentation. And the current state of the particle was marked as s_t^p, which contained the value of instantaneous location, dimensions and velocities of the individual particle.

2.1.4. State Estimation

The final state estimate was obtained by the average of the states obtained from each particle as shown in (4), i.e.

$$S_t = \frac{1}{N} \sum_{p=1}^{N} s_t^p \tag{4}$$

Also, the final combined likelihood parameter for the particle filter track at time t was found in similar manner and given by (5).

$$W_t = \frac{1}{N} \sum_{p=1}^{N} w_t^p \tag{5}$$

In the above equation, N denotes the total number of particles taken for tracking that were 50 for all the experimentation done in this research work. Whereas S_t, W_t stand for the combined weight and state estimate for the particle filter output for a single frame track procedure.

2.1.5. Resampling

Resampling process has been carried out to replace the particles that obtained lower likelihood with other particles in the particle sample for the next iteration. This process has been repeated in every frame to optimize the particle filter efficiency by filtering out the least required samples. The proposed algorithm exchanged the unworkable particles with S_t, which is the obtained mean state of the particles.

The individual track features were generated for the visible and thermal imagery counterpart separately and are denoted as W_t^V, S_t^V and W_t^T, S_t^T respectively. With the completion of the features extraction from the individual tracker, the final outcome was achieved using a new fusion technique, which has been presented in the following section of the paper.

2.2. Fusion Strategy

After obtaining the bimodal track information, selection of a suitable fusion rule was made to obtain track estimate. The fusion process mainly uses two critical statistics from the individual domain tracks to formulate its decision on the final track output estimate. The scheme, in general, can be illustrated using the following diagram in Fig. 1.

The fusion algorithm selects different fusion rules based upon the value of some observed parameters. First, the parameter ‘Difference in Weights DIW_t has been calculated (given by (6)). Based upon DIW value, it selected one of two cases.

$$DIW_t = |W_t^V - W_t^T| \tag{6}$$

If DIW_t value was found to be higher than a regularly updated threshold value wd_t, then the sensor mode having the more significant weight was chosen as the correct tracking source after validating it through a background-foreground similarity check. For every frame, the value of wd_t was automatically calculated using (7).

$$wd_t = \frac{1}{2}\left(\frac{W_t^V + W_t^T}{2}\right) \tag{7}$$

The above said validation process in turn, helps to rectify the decision in case of presence of camouflaged backgrounds. Had the higher weighing sensor been tracking deceptively due to disguising conditions, it would have shown high background similarity value and the result would have been discarded. The details of the implementation have been provided in section 2.2.1 of the paper.

In next case, if DIW_t was found to be less than wd_t, that suggests both domains have comparable confidence level in their output track estimate. Hence, an added criteria was used to obtain an error free decision. For overcoming this situation, the section procedure now looked upon for the spatial information, which was established based on founding a new parameter ‘Distance in States’, DIS_t. If (X_t^V, Y_t^V) and (X_t^T, Y_t^T) were the initial coordinates in 2D space of region of track outcomes in a single sensory domain, then DIS_t was found as the Euclidean distance between these points in image space and is defined in (8) as:

$$DIS_t = \sqrt{(X_t^V - X_t^T)^2 + (Y_t^V - Y_t^T)^2} \tag{8}$$

Further, the state difference threshold value was modelled to make spatial information based decisions. Value of space difference threshold was chosen in reference to dimensions of the object being tracked. Value of sd was found using (9) which helped in revealing whether the difference was significant or not.

$$sd = \frac{1}{2}\sqrt[2]{\left(\frac{A}{2}\right)^2 + \left(\frac{B}{2}\right)^2} \tag{9}$$

As mentioned earlier, A, B are the initial width and height of the object to be tracked.

Now, if value of DIS_t calculated was also minor, then it could be decided that estimated states were pointing towards a similar or nearby region in image space with alike confidence. Hence, both trackers resembled to an identical track estimate. So, the required output was found ready without using any Background validations that in turn saved lot of computational cost through a simple Foreground fusion framework detailed in section 2.2.2.

However, if DIS_t was significant (higher than value of), then it would have signified that although the weights obtained from individual trackers were similar, but these were indicating to two different locations. So, again background information was incorporated to aid the handicapped lone foreground detail. Section 2.2.3 describes the fusion for this case. The complete procedure of fusion strategy is as follows:

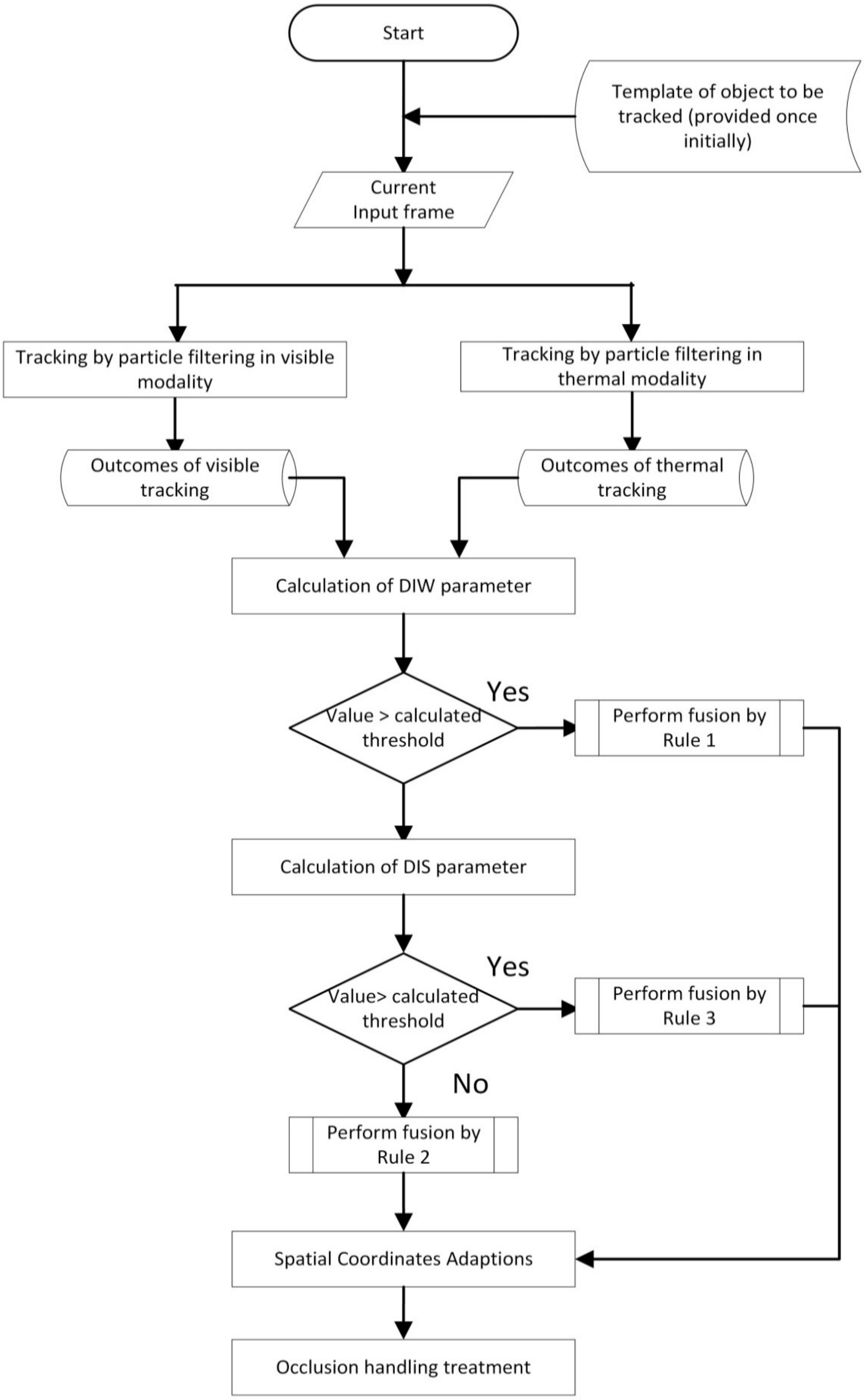

Start
Template of object to be tracked (provided once initially)
Current Input frame
Tracking by particle filtering in visible modality
Tracking by particle filtering in thermal modality
Outcomes of visible tracking
Outcomes of thermal tracking
Calculation of DIW parameter
Yes
Value > calculated threshold
Perform fusion by Rule 1
Calculation of DIS parameter
Yes
Value> calculated threshold
Perform fusion by Rule 3
No
Perform fusion by Rule 2
Spatial Coordinates Adaptions
Occlusion handling treatment

Fig. 1. Block representation of the work-flow procedure

2.2.1. Fusion Rule 1: $DIW_t > wd_t$ (Average weight obtained from one imaging modality was very high and obtained from other was too low).

In this case, due to very low likelihood indication given by one modality in self's decision, its opinion was not taken into consideration and the modality showing a superior likelihood of track was chosen as a sole leader for providing the track estimate. Further the sole dependency was authenticated only after passing the foreground–background validation process. Different techniques are available in the literature for finding the background region and foreground region relation. Our method adopted the local background model implementation of [17] to determine the background histogram of the state found.

If, the Bhattacharya similarity between reference target histogram and background region histogram found using (2) came out to be less than 0.5 (scale of 0-1), the following procedure for fusion was adopted:

i. Append that imagery's weight equal to unity, whose likelihood average weight was found to be far more than the other confidence diminished imagery.
ii. Make the other imagery's weight equal to zero.
iii. After changing the weights of individual imagery to either 0 or 1, the fusion output was calculated as:

$$S_t^F = \delta_t^V \times (S_t^V) + \delta_t^T \times (S_t^T) \tag{10}$$

Here δ_t^V and δ_t^T stand for modified weight values under the above said situation. The variable S_t^F denotes the output of fusion at time t after application of suitable fusion rule. The notation used for the fused output remains same throughout this paper.

Else, the situation was treated as no track in the current frame because, one tracker was having very low confidence in tracking and the other with a superior confidence was tracking wrongly either due to a deceiving condition or presence of similar background in its image signatures.

2.2.2. Fusion Rule 2: $DIW_t < wd_t$ and $DIS_t < sd$ (Average likelihood obtained from both imaging modalities and the location of track estimates from these were in close approximation).

Since the average confidence obtained from both imageries was nearby and also the output track locations pointed by each imagery resided closely in this case; therefore, a conjunctive consensus could be easily reached by use of a simple calculative procedure detailed in following steps:

i. Append Value of Foreground Adaptive Measure, γ, was found using (11) as:

$$\gamma = \frac{W_t^V}{W_t^V + W_t^T} \tag{11}$$

ii. The fused state estimate was updated according to (12).

$$S_t^F = \gamma \times (S_t^V) + (1 - \gamma) \times (S_t^T) \tag{12}$$

Hence, background calculations could be avoided for these situations that lessen the time consumption of the overall track process. The following subsection provides the implementation details of the last case of fusion.

2.2.3. Fusion Rule 3: $DIW_t < wd_t$ and $DIS_t > sd$ (Likelihood of both imaging modalities were found to be nearby but individual state estimate differed significantly).

For this different situation, the extra information that helped dissipate the ambiguity for the fusion was once again the local background conditions of the estimated track state outcome. The background details were found with the help of [17] as mentioned in subsection 2.2.1. Now, a reliable way to weigh the particles can be formulated in the following manner:

i. Value of adaptive background measure, β_t, was calculated based upon the similarities of backgrounds with the object reference template and was found using (13) as:

$$\beta_t = \frac{B_t^T}{B_t^T + B_t^V} \tag{13}$$

Where, B_t^T and B_t^V were the calculated Bhattacharyya similarity measures between the histograms of target object template and the background Regions of the states estimated in thermal and visible modality respectively. Both of these were found using (2).

ii. After obtaining a background measure, an effective way to provide confidence measures to the individual modes was adopted. This modelling is shown by (14) and (15):

$$\alpha_t = \frac{\beta_t . W_t^V}{\beta_t . W_t^V + (1-\beta_t) . W_t^T} \tag{14}$$

$$S_t^F = \alpha_t \times S_t^V + (1 - \alpha_t) \times S_t^T \tag{15}$$

Here, α_t is the parameter representing the weight adaptation parameter resulted from the introduction of the local background details of individual domain imagery. Rest of the parameters is the same as defined earlier in the paper.

The above rule indicates that the imaging modality that exhibited less similarity in its background region was multiplied with a larger weight factor as compared to the other one.

2.3. Improvement in Location of Track Estimate After Fusion

After updating the weights, it was necessary to cross-examine the relative spatial differences between the individual state estimates, as this could also degrade the performance of tracker. If one tracker is tracking correctly and has got a better confidence measure after above fusion process, but still the other tracker having even a small confidence measure could bring a significant shift in the correct track position. If it was pointing to a tracking estimate very far from the current position of the original target. So, despite adaptation in weights, the final track estimate might diverge without making an adaption in locations provided by individual trackers. Changes in states were made using the help of parameter relative to the weighting of individual trackers. First, W_{min} and W_{max} were found using (16) and (17) respectively.

$$W_{min} = \min (W_t^V, W_t^T) \tag{16}$$

$$W_{max} = \max (W_t^V, W_t^T) \tag{17}$$

After this, a factor was found for relative distance adjustments using the following relation presented in (18).

$$D = sd * \frac{W_{min}}{W_{max}} \tag{18}$$

After modification in weights, a check was made that the X and Y coordinates of the Track estimate region of least weighted domain differed (either less or more) from the X and Y coordinates of the other domain (having higher weight) by more than D pixels. If this condition existed, then a corresponding reduction or increment was made in the value of small weighted imagery's state coordinates so that their changed value had at most D pixels difference to remedy the unwanted shift. Otherwise, the values were kept unchanged.

2.4. Occlusion Rule

The method took the help of both the imaging modalities to detect the situation of occlusion in tracking. In the case of occlusion, the object or its parts hide behind another object present in the scene. So, the likelihood value suddenly gets decreased on the insertion of occluding. The proposed algorithm checked the similarity of the final track estimate in every frame with original track template available. This check was performed in both imagery domains concerned.

Moreover, the similarity measures obtained from both the domains were weighted with respect to their weights obtained after fusion. Now, the combined likeness degree was calculated using (19) as:

$$OS_t^F = W_t^V \times (OS_t^V) + W_t^T \times (OS_t^T) \tag{19}$$

Here, OS_t^V represents the Bhattacharyya measure between histogram densities for region of fused track estimate and the template histogram in the visible colour domain and OS_t^T is the corresponding counterpart in the thermal image. The values of these have been found using (19). If the value OS_t^F appeared to be higher than a threshold value for similarity, routine tracking process was continued. But if occlusion was detected due to a degraded similarity index, then algorithm moved towards occlusion handling procedure.

Once the occlusion was detected, then the algorithm searched for the robust presence of the object with the help of background subtraction using dynamic matrix factorization [27]. Once it was detected that the object comes out of occlusion, then the template model was updated by using the new conditions of the object, and the particles were redistributed around its location. The further routine tracking process continued till there is no occurrence of next occlude.

3. Experimental Results and Discussion

To verify the performance and scrutinize it with other states of the art trackers, we used MATLAB software on an I3 processor. The video sequences used along with the challenging conditions contained in these are written in the form of the following points:

i. OCTBVS Dataset Sequence 2 (Similarity in Backgrounds and visible camouflaging).
ii. OCTBVS Dataset Sequence 4 (Occlusion, Clutter)
iii. OCTBVS Dataset Sequence 6 (Presence of similar appearance objects, thermal camouflaging)
iv. INO Snow Parking Dataset (multiple alike objects, same appearance backgrounds, object rotation and most importantly effect of clutters).

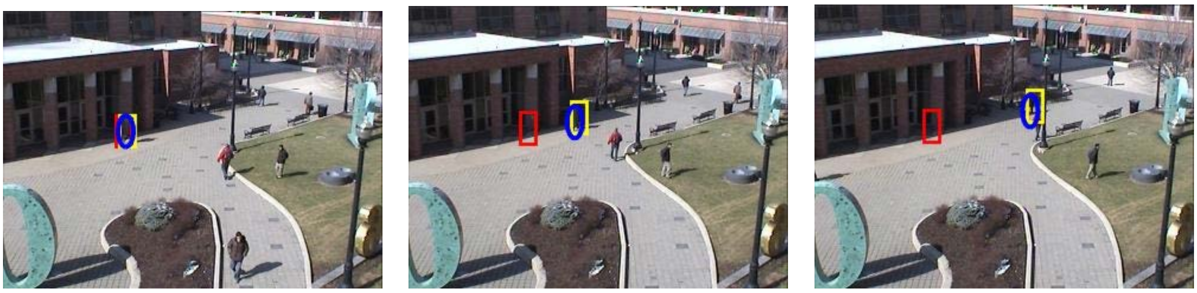
Experiments were performed and the results obtained were compared with 4 other algorithms viz. a) Colour Particle Filter, CPF [14], (b) Continuously adaptive data fusion of colour and gradients. CADF [18], (c) Adaptive fusion of thermal and Visible, AFTV [15], and (d) Tracking before fusion of infrared and visible, TBF [17]. The complete discussion on the results obtained has been narrated by observing the qualitative traits as well as the comparative statistical calculations from the performed experimentation.

3.1. Qualitative Demonstration of the utility of algorithm

The quality of a fusion algorithm can be judged from the fact that it should overcome the hurdles due to several conditions and unaffectedly chose the best outcome possible. Its performance should remain robust. This fact has been illustrated with the help of two video sequences:

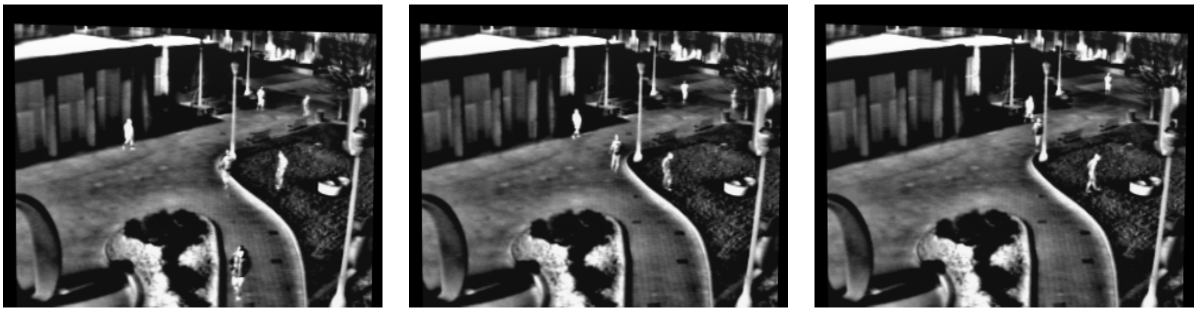
3.1.1. OCTBVS Data Sequence 2

The advantage of fusion of the thermal modality in aid to the visible tracking can be demonstrated by Fig. 2 and Fig. 3. Here, the object moves across a colour camouflaged background conditions. It can be seen that there are dark spaces between the pillars of the building and also a tree present on the right side of the view that is making difficult for target judgment in the visible domain.



(a) (b) (c)

Fig. 2. Track outcomes shown on visible image view of seq. 2. Individual colour based tracking (red rectangle), Individual thermal imagery tracking (yellow rectangle) and fusion based tracking using the proposed algorithm (blue ellipse). (a) Frame no. 16, (b) Frame no. 89, (c) Frame 193



(a) (b) (c)

Fig. 3. Thermal Image counterpart of Fig. 2 in the sequence 2 of OSU dataset. (a) Frame no. 16, (b) Frame no. 89, (c) Frame 193

Whereas the thermal signatures robustly mark the presence of the object as compared to the visible captures due to the temperature difference between the human object and the background. It is interesting to note that when both the trackers are tracking accurately, the algorithm does not indulge in any complex computations by avoiding the application of any testing criteria for decisions. However, as soon as the visible particle filter starts estimating an erroneous trajectory, then the final fused track estimate remains with the accurately tracking domain due to the re-arrangements of the weighing criteria and proposes fusion state adjustments.

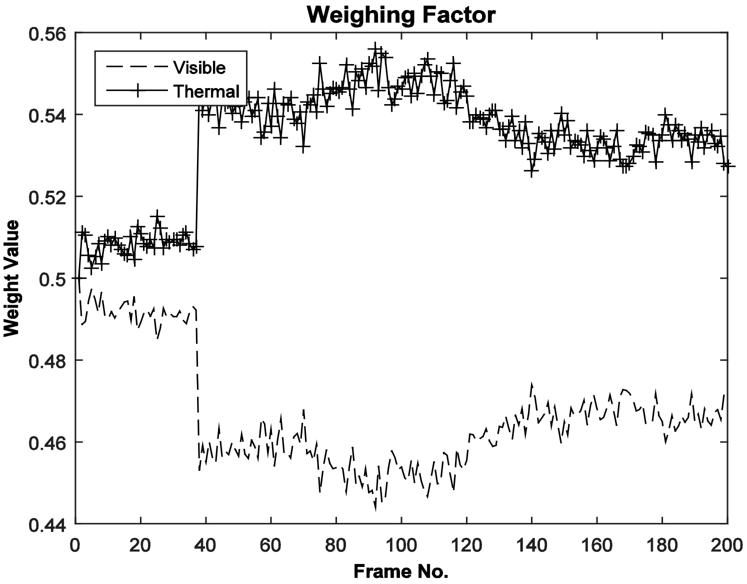
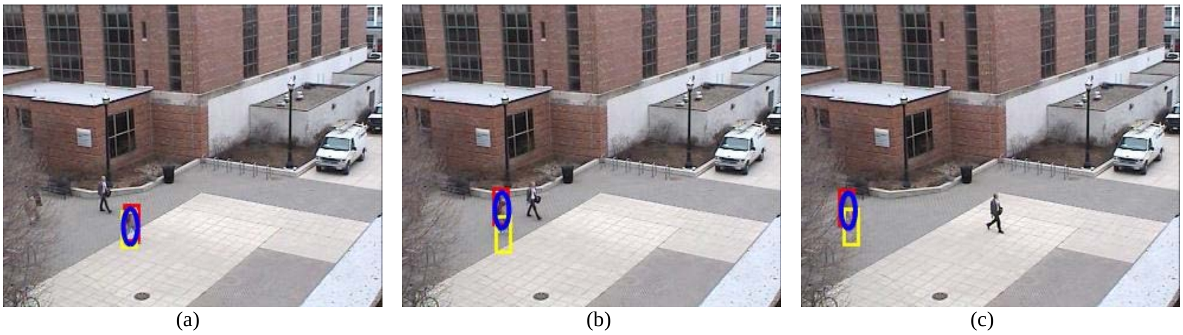


Fig. 4. Wight variation during sequence 2 videos

The fact can also be supported by Fig. 4, which shows that the weight of wrongly tracking visible domain starts decreasing and the fusion algorithm starts favouring the right tracking thermal domain after about 25 frames.

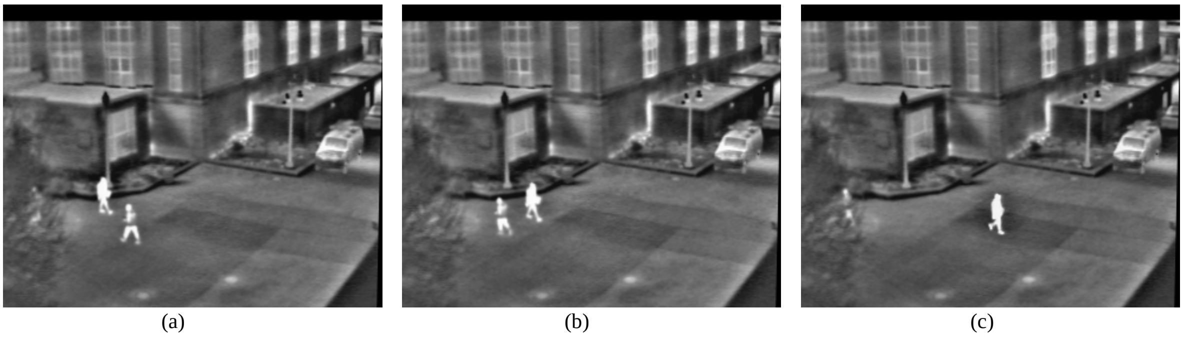
3.1.2. *OCTBVS Data Sequence 4*

The illustration shown in Fig. 5 and Fig. 6 suggests that the performance quality of the algorithm is not dependent solely on single imagery. The output during different frames of sequence 4 does not get affected when the thermal tracker tracks to incorrect output. Here rather than following the thermal imagery (which was earlier providing better estimates in the previous dataset), now the algorithm tracks by weighing in favour of the accurate tracking visible imagery. So, the fusion approach depreciates the role of thermal imagery in case of it providing false estimates.



(a) (b) (c)

Fig. 5. Track outcomes shown on visible image view of seq. 4. Individual colour based tracking (red rectangle), Individual thermal imagery tracking (yellow rectangle) and fusion based tracking using the proposed algorithm (blue ellipse). (a) Frame no. 45, (b) Frame no. 81, (c) Frame 153



(a) (b) (c)

Fig. 6. Thermal Image counterpart of Fig. 5 in the sequence 2 of OSU dataset. (a) Frame no. 45, (b) Frame no. 81, (c) Frame 153

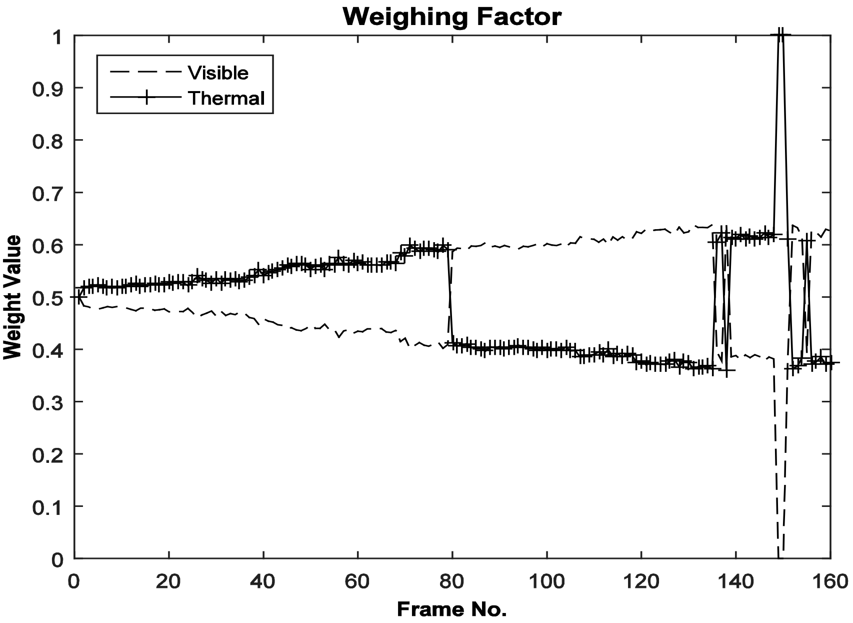


Fig. 7. Weight variation during sequence 4 videos

The same can also be illustrated through Fig. 7, which shows weight values provided to individual imageries after the fusion process. It is interesting to mention that under normal conditions till frame no. 78, the proposed technique finds the track estimate using simple weighted fusion as mentioned by fusion rule number 2 in the paper. This estimation used simple calculations. However, when a new person appeared near to the tracked target, there appeared some difference in opinions of individual trackers. Consequently, the algorithm shifted to fusion rule 3 and started to adjust weights more in favour of accurately tracking visible imagery.

3.2. Quantitative Comparisons with other methods

The accuracy was judged while comparing the experimental results with the ground truth positions of the object throughout the video sequence. This comparison was quantified in terms of two parameters Root Square Tracking Error and the statistical F-measure calculated with the help of precision and recall values. The significance of these parameters is that the value of one resembles the closeness in the position of the target tracked with the actual position of the object and the other represents the bounding box overlap efficiency of the results. The combined diversion from the ground truth positional values in both axes was represented in the form of Root Square Tracking Error, $RSTE_t$, whose value can be found using (20).

$$RSTE_t = \sqrt{\left(G_t(x) - S_t^F(x)\right)^2 + \left(G_t(y) - S_t^F(y)\right)^2} \tag{20}$$

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

The next parameter F-measure is found by first finding the precision and recall values and then substituting these in (23). The precision (found using (22)) is defined as the ratio of an area covering intersection the ground truth bounding box with the estimated track bounding box to the area covered by the bounding box of the estimated track state. Whereas, recall can be found using (21), i.e., by the ratio of an area covering intersection the ground truth bounding box with the estimated track bounding box to the area covered by the original bounding box enclosing the object.

$$R_t = G_t(area) \cap S_t^F(area)/G_t(area)\;) \tag{21}$$

$$P_t = G_t(area) \cap S_t^F(area)/S_t^F(area) \tag{22}$$

$$F_t = \;2\; \times P_t \;\times R_t/(P_t + R_t) \tag{23}$$

Here, P_t and G_t mark the current frame precision and recall values respectively. Whereas F_t is the statistical F-measure of the obtained track estimate.

3.2.1. OCTBVS Data Sequence 2

This sequence of OSU dataset contains a human pedestrian moving past a colour similarity background. It has been observed that as soon as the object passes the building walls after frame number 60 and is left only with black background (similar to the colour of dress the pedestrian is wearing), the ordinary colour-based particle filter turns out to be misguided and remains attached to the black background instead of tracking the real target. This effect is illustrated by the high deviation in the error curve shown in Fig. 8

Also, the continuously adaptive data fusion model of gradient and colour fusion remains in the proper track search till mid-way but loses the actual trajectory as it starts weighing more in favour of the colour domain which is already providing false track. Now, coming to the approach of thermal and visible fusion in [17], the track estimates are at some distance apart from the original position of the object consistently. This situation arises because though this fusion method improves the weights for the tracker by weighing more in favour of the thermal imagery, it does not take care of the positional coordinate arrangement. The larger the offshoot distance of false sensing domain, more deviation is present as observed in Fig. 8 and Fig. 9.

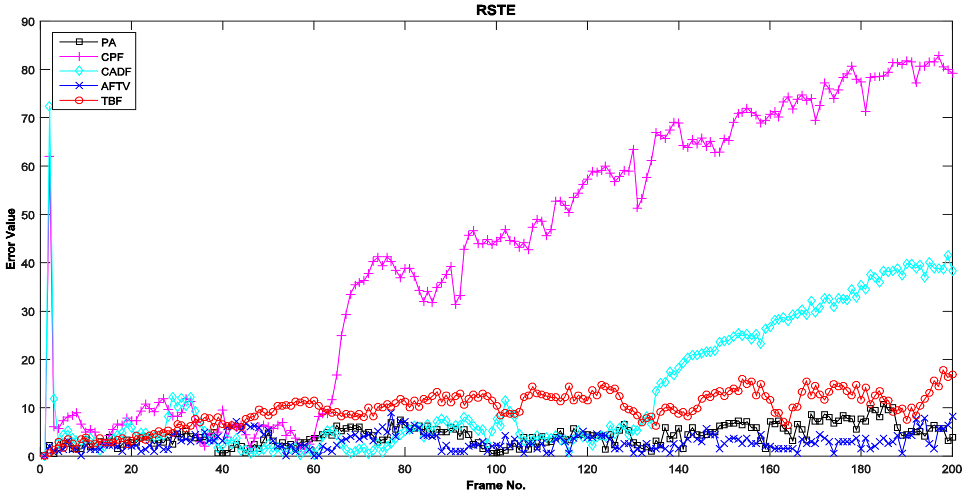


Fig. 8. RSTE plot for Sequence2

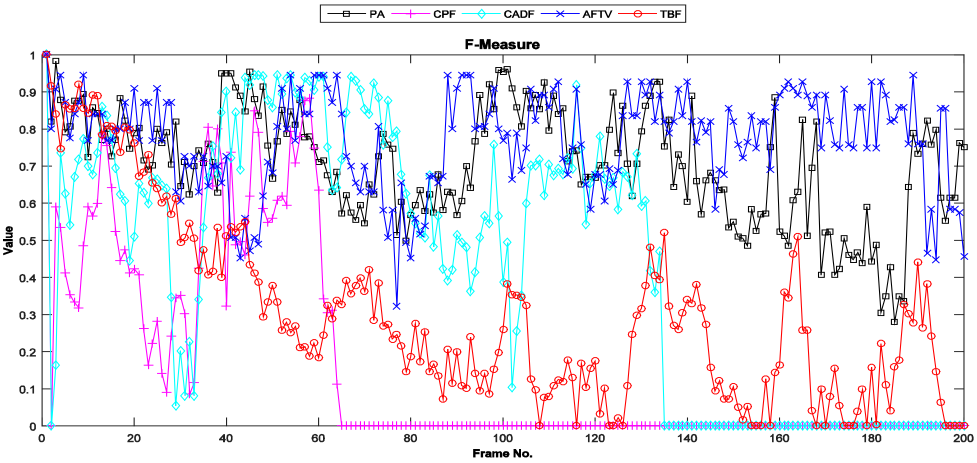


Fig. 9. F- Measure plot for Sequence2

The adaptive fusion of thermal and visible domain by [15] is providing correct weighting by adaptive background fusion model to weigh in favour of the thermal domain signatures of the view. However, this method is not computationally effective as it calculates a background model for every particle in every frame and then performs a similarity comparison for each of them. On the other side, the proposed method takes care of the track estimate by providing parameters both for weights updation as well as track state position updation. As indicated by Fig. 8 and Fig. 9, although the particle filter in colour domain is tracking falsely after frame 60, the thermal domain tracker keeps providing accurate track outcome. Accordingly, the proposed algorithm starts weighing more in favour of thermal domain tracker. Besides, a more substantial drift in location by the result of a false tracking domain has been encountered by the algorithm.

3.2.2. INO Dataset Snow Parking Sequence

The tracking in the snow parking dataset contained many challenging situations such as the presence of multiple alike objects, same colour backgrounds, object rotation and severe effect of clutters in the scene inside both imagery domains. The illustration is provided in Fig. 10 and Fig. 11.

It is notable to see that few other algorithms tend to lose their way as the moving target stops on the path for few frames and it interacts with a similar appearance pedestrian and rotates his motion direction several times during his communication. After frame 770, the object comes across a car that acts as a similar background object. Here rest of all trackers except the individual colour-based particle filter track correctly. Near to frame 780, the target comes in contact with a similar pedestrian that acts as clutter in both imageries, and hence many algorithms tend to hinder away from the original object.

This hazard has been indicated by an increase in the RSTE value (Fig. 12) and a decrease in the F-Measure value (Fig. 13). The AFTV algorithm also suffers ambiguities in the track estimate during frame interval of 790-860, under which the object rotates several times in various directions and comes back to original motion direction only after frame 870.

The TBF algorithm suffers during the presence of clutter and starts tracking correctly afterwards. The proposed algorithm performs a reasonable job throughout the whole frame interval due to amendments made both in weighing factors and the spatial positions.

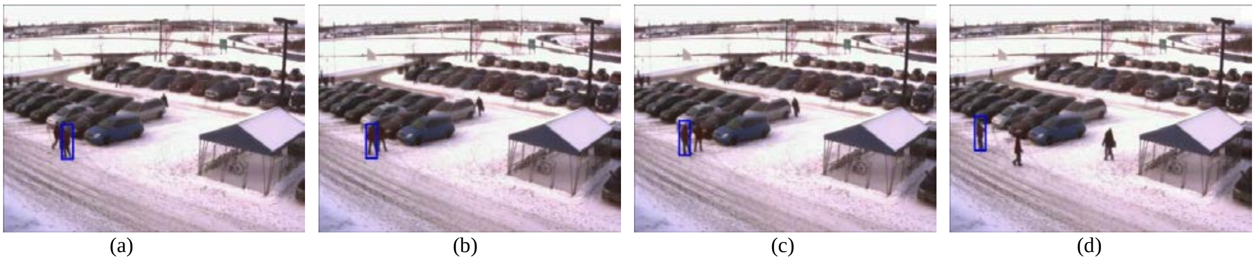

(a) (b) (c) (d)

Fig. 10. Track outcomes of proposed algorithm shown on visible image view of SPDS video sequence (a) Frame no. 782, (b) Frame no. 796, (c) Frame 811, (d) Frame no.880

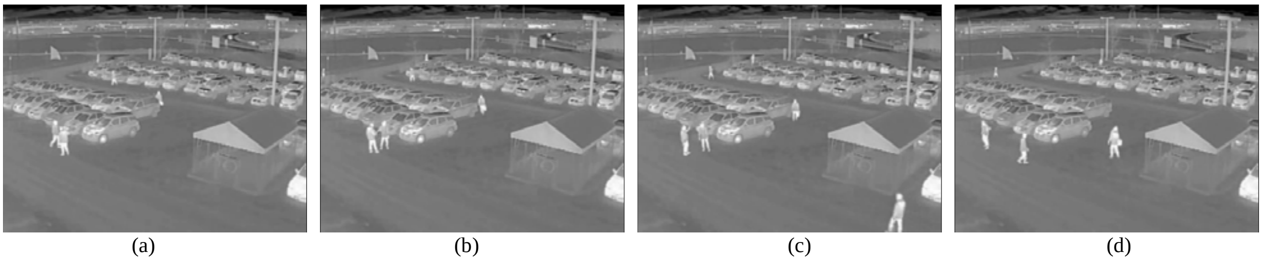

(a) (b) (c) (d)

Fig. 11. Thermal Image counterpart of Fig.10 in the sequence 2 of OSU dataset. (a) Frame no. 28, (b) Frame no. 60, (c) Frame 88, (d) Frame no.145

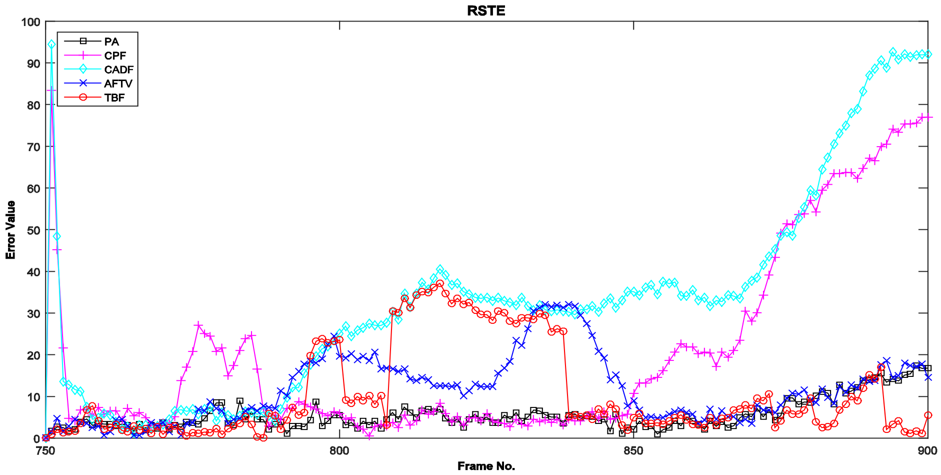


Fig. 12. RSTE plot for SPDS

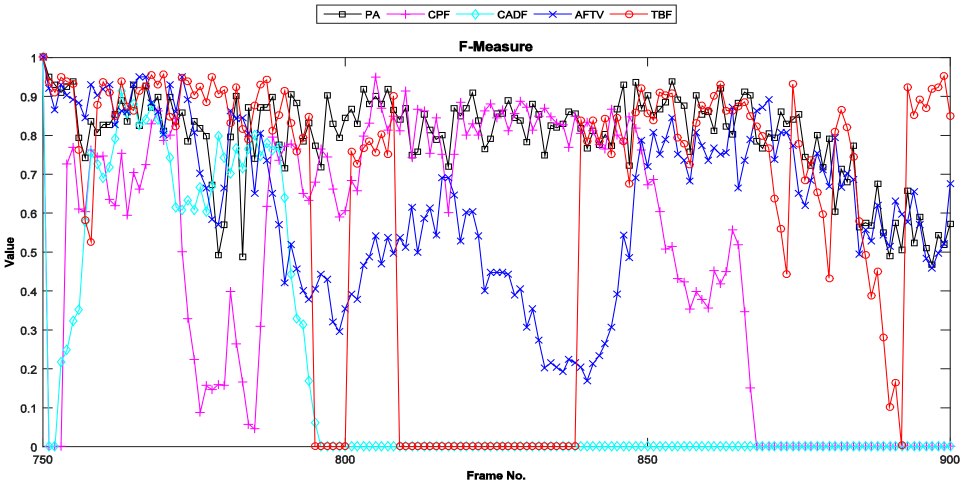


Fig. 13. F-Measure plot for SPDS

3.2.3. *Sequence 6 Dataset*

The scenes of this dataset contain challenging problems for a visual tracker like changing background conditions, the presence of similar feature objects, and also a similarity in the thermal domain. The object is initially in front of a white coloured car with black bottom front, and a black appearing mirror, but then its motion is traversed against empty space followed by passing a similar coloured pillar that also has it's lower half-size comparable to the dimensions of the object to be tracked. This situation engraves a challenging task, which is shown in Fig. 14 and Fig. 15.

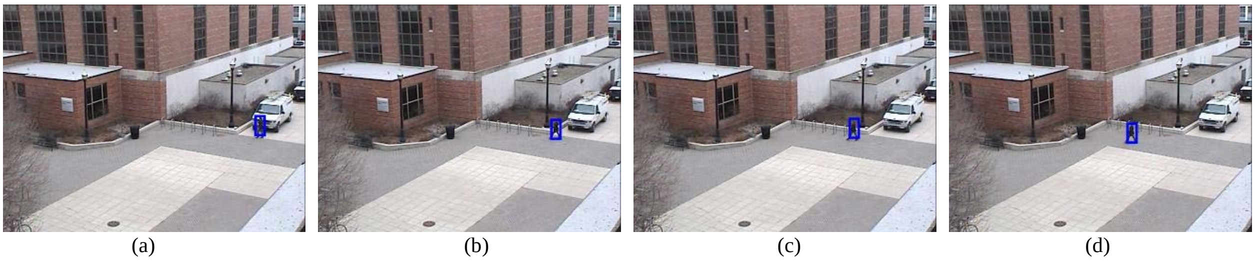

(a) (b) (c) (d)

Fig. 14. Track outcomes of proposed algorithm shown on visible image of seq. 6. (a) Frame no. 28, (b) Frame no. 60, (c) Frame 88, (d) Frame no.145

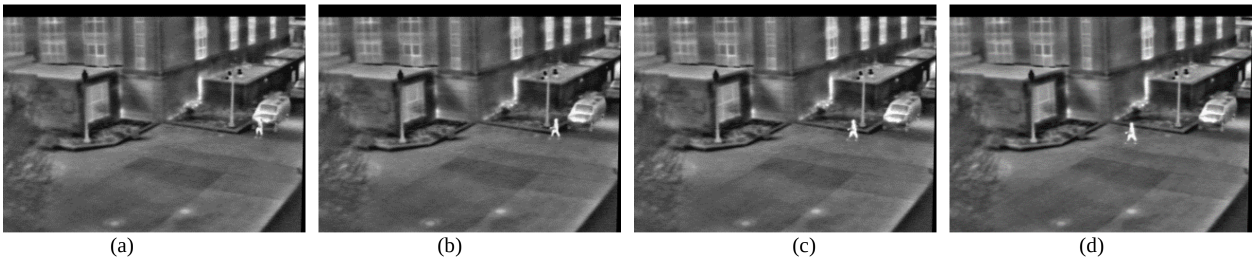

(a) (b) (c) (d)

Fig. 15. Thermal Image counterpart of Fig.14 in the sequence 6 of OSU dataset. (a) Frame no. 28, (b) Frame no. 60, (c) Frame 88, (d) Frame no.145

Further, the presence of various coloured clutter objects such as black dustbin, windows of the building, etc. makes it tough for tracking using a lonely colour based tracker. The thermal image of the scene also faces a background camouflage in the initial part of the sequence that can be seen through Fig. 14(a).

No trackers except for two algorithms produced an accurate track estimate for this dataset containing various difficulties. The correct tracking was only achieved till nearby frame no. 40. The CADF was working based on the fusion of HOG and colour feature. However, it favoured the colour cue (not tracking accurately) as illustrated by the CPF curve in Fig. 15 due to more significant deviations in the colour likelihood of particle filter. The resultant is the start of ramp increase in the RTSE and decrease in F-Measure value (Fig. 16, Fig. 17). Though TBF favours mainly in the thermal domain from the start, it tracks below par due to thermal camouflaging present the individual thermal tracker. Moreover, with no adjustment in the final state positions, the fusion deviates a lot from the actual position of the target.

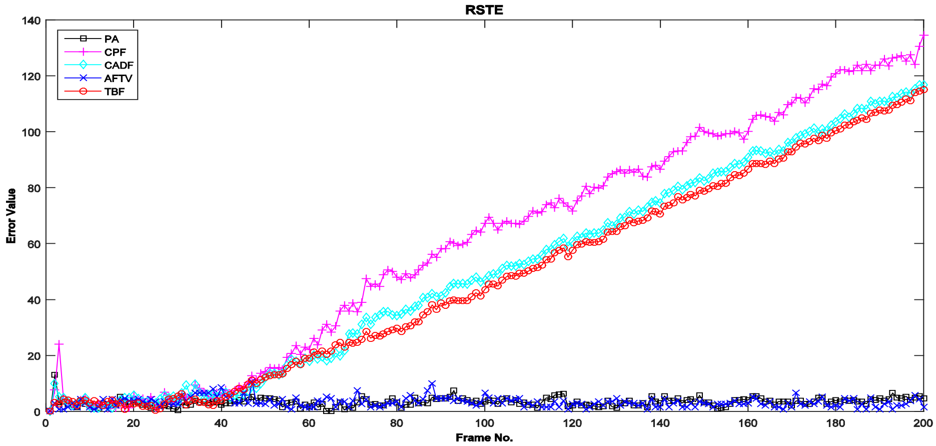


Fig. 16. RSTE Plot for Sequence 6 video

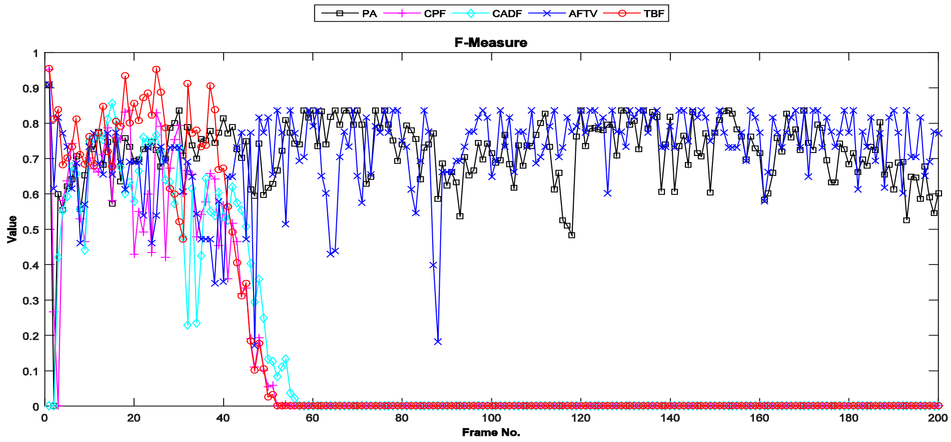


Fig. 17. F- Measure Plot for Sequence 6 video

The results have also been compared via a tabular representation providing a clear view of the performance of the different algorithms. Table 2 provides the time to process analysis, whereas Table 3 presents the RSTE comparison and F-measure statistics have been illustrated using Table 4. It is clear from Table 2 that the proposed algorithm only lacks behind the basic particle filter in timing analysis, but takes less time in comparison to other fusion schemes.

RSTE for various algorithms value was found using (23) for each frame. The average error has been tabulated in Table 3. All the results are averaged over five runs of experiments due to the random nature of the particle filter incorporated in each method. The last column of the table shows the overall average performance.

The best has been written in bold text and italics has been used for the second position. It is interesting to note that though the proposed algorithm does not top the charts for every dataset, it maintains a constant healthy position throughout all the datasets containing diverse, challenging conditions. This performance has been shown by the overall best low RSTE value and largest F-Measure values. The AFTV also performs relatively well, but it requires many computations due to the performance of tedious fusion process for no. of times in a single frame track estimate. This fact is supported by a high time to process details provided in Table 2. Among rest of algorithms, though some perform fast (CPF TBF) but lack in the accuracy counterpart. These provide good results only for a few sequences but perform poorly for other videos. Hence, their overall RSTE and F-Measure in Table 3 and Table 4 are not appreciable respectively.

Table 2. Time to Process (in sec/frame)

Algorithm used	Video Sequence Used				
	Seq. 2	Seq. 4	Seq. 6	Snow Parking	Average
PA	*.25*	*.23*	*.20*	*.30*	*.25*
CPF	**.22**	**.22**	**.15**	**.29**	**.22**
CADF	.59	.56	.43	1.1	.67
AFTV	.90	.80	.65	1.3	.91
TBF	.25	.26	.22	.39	.28

Table 3. The RSTE values

Algorithm used	Video Sequence Used				
	Seq. 2	Seq. 4	Seq. 6	Snow Parking	Average
PA	*.71*	**.78**	**.77**	**.81**	**.77**
CPF	.16	.69	.13	.51	.37
CADF	.44	.70	.14	.19	.37
AFTV	**.77**	.66	**.77**	*.63*	*.71*
TBF	.30	*.72*	*.17*	.61	.45

Table 4. The F-Measure values

Algorithm used	Video Sequence Used				
	Seq. 2	Seq. 4	Seq. 6	Snow Parking	Average
PA	*.71*	**.78**	**.77**	**.81**	**.77**
CPF	.16	.69	.13	.51	.37
CADF	.44	.70	.14	.19	.37
AFTV	**.77**	.66	**.77**	*.63*	*.71*
TBF	.30	*.72*	*.17*	.61	.45

4. Conclusion

In the proposed algorithm, a unique way of tracking by the combined use of visible imaging and thermal imaging was presented. The method formulated a decisive fusion strategy by analyzing the proximity between parameters found from both the imaging modalities. The parameters used for developing the fusion strategy comprised of the average likelihood weight of particles (obtained from individual imagery) and the location value in image space pointed by the track output of individual imaging tracker. It is clear from the experiments that the presented algorithm performed efficiently under various challenging conditions. It provided commendable accuracy along with consuming small processing time as compared with the state-of-the-art methods. The proposed method top scored in case of two sequences and a second position in the rest two data sequences taken during experiments when root square tracking error was found and compared with different state of the art methods. Further the overlapping efficiency in terms of F-Measure was best for three sequences out of four. The proposed method fared with overall RSTE of 4.77 and an F-Measure of 0.77 when averaged over various data sequences. Time consumed was also less in comparison to state of art fusion based trackers and lagged behind only the individual colour image based particle filter. In addition, the fusion scheme also appears to have the advantage over other techniques in term of the fact that it is independent of the tracking algorithm chosen at the individual image sensor level. It requires only two parameters from the output of any tracker: i) Its track confidence measure and, ii) the position of the output track estimate of the tracker. Both of these traits are readily available for an individual tracker. In future, the method can be applied to tracking under bimodal fusion of imageries through any tracker, but more work needs to be done in the direction to formulate scaled versions of such fusion techniques that can be applied more than two modalities.

References

[1] Ali, U., Ali, M.: ‘Optimized Visual and Thermal Image Fusion for Efficient Face Recognition’ in ‘9th Int. Conf. on Inf. Fusion’, (2006), pp. 1–6.

[2] G. Bebis, A. Gyaourova, S.S. and I.P.: ‘Face Recognition by Fusing Thermal Infrared and Visible Imagery’ Image Vis. Comput., 2006, 24, (7), pp. 727–742.

[3] Heo, J., Kong, S.G., Abidi, B.R., Abidi, M.A.: ‘Fusion of visual and thermal signatures with eyeglass removal for robust face recognition’, in ‘IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops’, USA, (2004).

[4] Kong, S.G., Heo, J., Boughorbel, F., et al.: ‘Multiscale fusion of visible and thermal IR images for illumination- invariant face recognition’Int. J. Comput. Vis., 2007, 71, (2), pp. 215–233.

[5] Wilhelm, T., Böhme, H.J., Gross, H.M.: ‘A multi-modal system for tracking and analyzing faces on a mobile robot’Rob. Auton. Syst., 2004, 48, (1), pp. 31–40.

[6] Cielniak, G., Duckett, T.: ‘Active People Recognition using Thermal and Grey Images on a Mobile Security Robot’ IEEE/RSJ Int. Conf. Intell. Robot. Syst. IROS, (2005), pp. 3610–3615.

[7] Cielniak, G., Duckett, T., Lilienthal, A.J.: ‘Improved data association and occlusion handling for vision-based people tracking by mobile robots’, in ‘2007 IEEE/RSJ International Conference on Intelligent Robots and Systems’ (2007), pp. 3436–3441

[8] Palmerini, G.B., Università, S.: ‘Combining Thermal and Visual Imaging in Spacecraft Proximity Operations’2014, 2014, (December), pp. 10–12.

[9] Tong, Y., Liu, L., Zhao, M., Chen, J., Li, H.: ‘Adaptive fusion algorithm of heterogeneous sensor networks under different illumination conditions’Signal Processing, 2016, 126, pp. 149–158.

[10] Zhou, Z., Wang, B., Li, S., Dong, M.: ‘Perceptual fusion of infrared and visible images through a hybrid multi-scale decomposition with Gaussian and bilateral filters’Inf. Fusion, 2016, 30, pp. 15–26.

[11] Ma, J., Ma, Y., Li, C.: ‘Infrared and visible image fusion methods and applications: A survey’Inf. Fusion, 2018, 45, (December 2017), pp. 153–178.

[12] Walia, G.S., Kapoor, R.: ‘Recent advances on multicue object tracking: a survey’Artif. Intell. Rev., 2016, 46, (1), pp. 15821–15847.

[13] Li, C., Wu, X., Zhao, N., Cao, X., Tang, J.: ‘Fusing two-stream convolutional neural networks for RGB-T object tracking’Neurocomputing, 2017.
[14] Nummiaro, K., Koller-Meier, E., Van Gool, L.: ‘An adaptive color-based particle filter’Image Vis. Comput., 2003, 21, (1), pp. 99–110.
[15] Talha, M., Stolkin, R.: ‘Particle filter tracking of camouflaged targets by adaptive fusion of thermal and visible spectra camera data’IEEE Sens. J., 2014, 14, (1), pp. 159–166.
[16] Conaire, C.Ó., O’Connor, N.E., Smeaton, A.: ‘Thermo-visual feature fusion for object tracking using multiple spatiogram trackers’Mach. Vis. Appl., 2008, 19, (5–6), pp. 483–494.
[17] Xiao, G., Yun, X., Wu, J.: ‘A new tracking approach for visible and infrared sequences based on tracking-before-fusion’Int. J. Dyn. Control, 2016, 4, (1), pp. 40–51.
[18] Xiao, J., Stolkin, R., Oussalah, M., Leonardis, A.: ‘Continuously Adaptive Data Fusion and Model Relearning for Particle Filter Tracking With Multiple Features’IEEE Sens. J., 2016, 16, (8), pp. 2639–2649.
[19] Stolkin, R., Rees, D., Talha, M., Florescu, I.: ‘Bayesian fusion of thermal and visible spectra camera data for region based tracking with rapid background adaptation’IEEE Int. Conf. Multisens. Fusion Integr. Intell. Syst., 2012, pp. 192–199.
[20] Fendri, E., Boukhriss, R.R., Hammami, M.: ‘Fusion of thermal infrared and visible spectra for robust moving object detection’Pattern Anal. Appl., 2017, 20, (4), pp. 907–926.
[21] Singh, S. Khosla, A., and Kapoor, R. ‘Visual-Thermal Fusion Based Object Tracking via a Granular Computing Backed Particle Filtering’, IETE Journal of Research, 2022, pp. 1-16.
[22] Research, S.R., Chan, A.L.: ‘Enhanced target tracking through infrared-visible image fusion’14th Int. Conf. Inf. Fusion, 2011, pp. 1–8.
[23] Singh, S. Khosla, A., and Kapoor, R., “Object Tracking with a Novel Visual-Thermal Sensor Fusion Method in Template Matching”, International Journal of Image, Graphics and Signal Processing, 2019, 11(7), pp. 39-47.
[24] Wu, Y., Blasch, E., Chen, G., Bai, L., Ling, H.: ‘Multiple source data fusion via sparse representation for robust visual tracking’, in ‘14th International Conference on Information Fusion’ (2011), pp. 1–8
[25] Li, C., Cheng, H., Hu, S., Liu, X., Tang, J., Lin, L.: ‘Learning Collaborative Sparse Representation for Grayscale-Thermal Tracking’IEEE Trans. Image Process., 2016, 25, (12), pp. 5743–5756.
[26] Li, C., Hu, S., Gao, S., Tang, J.: ‘Real-time grayscale-thermal tracking via laplacian sparse representation’, in ‘Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)’ (2016), pp. 54–65
[27] Liu, H.P., Sun, F.C.: ‘Fusion tracking in color and infrared images using joint sparse representation’Sci. China-Information Sci., 2012, 55, (3), pp. 590–599.
[28] Yong, H., Meng, D., Zuo, W., Zhang, L.: ‘Robust Online Matrix Factorization for Dynamic Background Subtraction’IEEE Trans. Pattern Anal. Mach. Intell., 2017.

Authors’ Profiles

Satbir Singh obtained his Ph.D. degree from the National Institute of Technology, Jalandhar (NITJ), and received M.E. in Electronics and Communication Engineering from Thapar University, Patiala, India. He is presently working with Centre for Artificial Intelligence, NITJ. Previously, he had worked with Central Scientific Instruments Organization - India, Delhi Technological, University, Delhi - India, and Centre of Advanced Computing, Mohali - India. His research interests include computer vision, artificial intelligence, image and signal processing, and IoT.

Arun Khosla received his Ph.D. degree from Indraprastha University, Delhi in 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 a reviewer for various IEEE and other National and International conferences and serves on the editorial board of the International Journal of Swarm Intelligence Research. He is a life member of Indian Society of Technical Education.

Rajiv Kapoor received an M.E. and Ph.D. degree in ECE from Delhi College of Engineering, Delhi University, and Punjab University, Chandigarh, respectively. Dr. Kapoor is presently working as Professor in Electronics Communication Engineering Department, Delhi Technological University, Delhi). He has authored over 100 research papers in various renowned international journals 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, "Object tracking via a Novel Parametric Decisions based RGB-Thermal Fusion", International Journal of Image, Graphics and Signal Processing(IJIGSP), Vol.15, No.4, pp. 1-18, 2023. DOI:10.5815/ijigsp.2023.04.01