Fusion of Thermal Infrared and Visible Spectrum Video for Robust Surveillance

Praveen Kumar, Ankush Mittal, and Padam Kumar

Department of Electronics and Computer Engineering,
Indian Institute of Technology, Roorkee,
India 247667
praveen.kverma@gmail.com, {ankumfec, padamfec}@iitr.ernet.in

Abstract. This paper presents an approach of fusing the information provided by visible spectrum video with that of thermal infrared video to tackle video processing challenges such as object detection and tracking for increasing the performance and robustness of the surveillance system. An enhanced object detection strategy using gradient information along with background subtraction is implemented with efficient fusion based approach to handle typical problems in both the domains. An intelligent fusion approach using Fuzzy logic and Kalman filtering technique is proposed to track objects and obtain fused estimate according to the reliability of the sensors. Appropriate measurement parameters are identified to determine the measurement accuracy of each sensor. Experimental results are shown on some typical scenarios of detection and tracking of pedestrians.

1 Introduction

With the advances in sensor and computing technologies, new generation video surveillance and security system will be required to be *persistent* (ability to function continuously for 24 hours and in a variety of scenarios) and *intelligent* in combining multimedia information for robust operation. Color and grayscale video cameras have an obvious limitation of daytime operation only, whereas Infrared media are more informative in dark environment (especially in night). Traditional approaches analyze video only in a single modality; either using the visible spectrum or using another modality such as mid-wave or long-wave infrared images [1]. Since, the visible spectrum and thermal infrared are inherently complementary; having their own unique characteristics, combining them can be advantageous in many scenarios, when either may perform poorly. For example, in foggy weather condition and in night, IR sensor will outperform visible range camera. Sudden lighting changes, shadows and camouflage, in visible spectrum, can often cause the foreground detection to incorrectly classify pixels. Combining the visible analysis with infrared imaging seems very beneficial, as it is very robust to the above mentioned problems. However, in good lighting and stable background conditions, visible spectrum video would give better results because of containing strong edges, robust color and other features with comparatively low noise. Although humans and other hot objects usually appear as areas

P. Kalra and S. Peleg (Eds.): ICVGIP 2006, LNCS 4338, pp. 528 – 539, 2006.

of high contrast and are more distinctive in infrared but well insulated clothing can cause the torso to have very low contrast and appear as background noise. Sudden temperature change, heat diffusion through objects in contact and “Halo effect” produced by some infrared sensors, which appears as a dark or bright halo surrounding very hot or cold objects respectively, are some additional difficulties that cause incorrect segmentation of object region. The challenge therefore is to determine the best approach to combine both modalities so that typical problems in both the domains can be addressed. This is made more challenging by the fact that some sources of data may give misleading or incorrect information. For example, changes in lighting, such as those caused by clouds blocking the sun’s light during the daytime, can cause incorrect change detection in the visible spectrum. In a recent review on surveillance research [2], Hu et al. conclude in their section on Future Developments in Surveillance that “*Surveillance using multiple different sensors seems to be a very interesting subject. The main problem is how to make use of their respective merits and fuse information from such kinds of sensors”*. In another review of video surveillance and sensor networks research [3], Cucchiara argues that the integration of video technology with sensors and other media streams will constitute the fundamental infrastructure for new generations of multimedia surveillance systems.

This paper presents an approach of fusing the information provided by thermal infrared video and that of visible spectrum video for robust object detection and accurate object tracking thereby increasing the performance and robustness of the surveillance system. An enhanced object detection strategy is implemented with efficient fusion based approach. We collected a database of known scenarios in indoor and outdoor situations captured simultaneously by video and IR cameras. These image sequences (video and IR) are time synchronized and geometrically corrected to co-register them with their counterparts. For both sensor sequences, we apply our enhanced background subtraction algorithm using gradient information, to identify region of interests (ROI) and extract blobs corresponding to the objects in the scene. For individual sensor sequence, blobs have to be matched with the objects (tracked at fusion level) present in the previous frame and some measurement parameters are computed. For tracking purpose, track-to-track fusion scheme is used, where a separate Kalman Filter is used for each track to obtain a filtered estimate. An intelligent fusion algorithm subsequently proceeds to obtain fused measurement data for each object according to the reliability of the sensors. A Fuzzy Inference System (FIS) is employed to assign suitable weights to each sensors filtered estimate, based on the value of two parameters called ‘*Confidence*’ and ‘*Appearance Ratio’,* computed for all the objects in each sensor output. Finally, a defuzzificator obtains the fused estimated measurement based on the weightage values. The Experimental results are done to demonstrate the effectiveness of fusing visible and IR in some typical scenarios of detection and tracking of pedestrians.

2 Literature Review and Background

Object Detection and Tracking beyond Visible Spectrum: Recent literature on the exploitation of near-infrared information to track humans generally deals only with the face of observed people and a few are concerned with the whole body [4] but these approach rely on the highly limiting assumption that the person region always has a much brighter (hotter) appearance than the background. In [5], the author proposes a novel contour based background subtraction strategy to detect people in thermal imagery, which is robust across a wide range of environmental conditions. First of all, a standard background-subtraction technique is used to identify local region-of interest (ROI), each containing the person and surrounding thermal halo. The foreground and background gradient information within each region are then combined into a contour saliency map (highlighting the person boundary). Using a watershed-based algorithm, the gradients are thinned and thresholded into contour fragments and A* search algorithm is used to connect any contour gaps. However use of highly computational techniques, makes their approach inappropriate for use in real time surveillance settings.

Modality Fusion: Multi modal fusion is the process of combining data from multiple sources (of different spectrum) such that the resulting entity or decision is in some sense better than that provided by any of the individual sources [6]. Data fusion techniques have had a long history in radar and vision based military applications to enhance the information content of the scene by combining multispectral images in one image. However, only recently data fusion is being considered for enhancing the capabilities of automatic video-based detection and tracking system for surveillance purpose. In [7], the fusion of thermal infrared with visible spectrum video, in the context of surveillance and security, is done at the object level. Detection and tracking of blobs (regions) are performed separately in the visible and thermal modality. An object is made up of one or more blobs, which are inherited or removed as time passes. Correspondences are obtained between objects in each modality, forming a master-slave relationship, so that the master (the object with the better detection or confidence) assists the tracking of the slave in the other modality. In a recent work, Davis et al. [8] propose a new contour-based background-subtraction technique using thermal and visible imagery for persistent object detection in urban settings. Their algorithm requires co-registered image from two streams. Statistical background subtraction in the thermal domain is used to identify the initial regions-of-interest. Color and intensity information are used within these areas to obtain the corresponding regions of-interest in the visible domain. Within each image region (thermal and visible treated independently), the input and background gradient information are combined as to highlight only the boundaries of the foreground object. The boundaries are then thinned and thresholded to form binary contour fragments. Contour fragments belonging to corresponding regions in the thermal and visible domains are then fused using the combined input gradient information from both sensors.

Multi-Sensor Fusion: The determination of the target's position and velocity from a noisy time-series of measurements constitute a classical statistical estimation problem and it involves the use of sequential estimation techniques such as the Kalman filter or its variants. Observational data may be combined, or fused, at a variety of levels from the raw data (or observation) level to feature level, or at the decision level. In the fusion process, it is essential to asses the reliability of sensor data because the results could be seriously affected in the case of malfunctioning sensor. Therefore for fusing data collected from different sensors requires the determination of measurements' accuracy so that they can be fused in a weighted manner. In [9], the authors propose a multi-sensor data fusion method for video surveillance, and demonstrated the results by using optical and infrared sensors. The measurements coming from different sensors were weighted by adjusting measurement error covariance matrix by a metric called Appearance Ratio (AR), whose value is proportional to the strength of the segmented blobs. In [10], the authors propose a hybrid multi-sensor data fusion architecture using Kalman filtering and fuzzy logic techniques. They feed the measurement coming from each sensor to separate fuzzy–adaptive kalman filters (FKF), working in parallel. Based on the value of a variable called Degree of Matching (D0M) and the measurement noise covariance matrix R coming from each FKF, a fuzzy inference system (FIS) assigns a degree of confidence to each one of the FKFs output. Finally, a defuzzificator obtains the fused estimated measurement based on the confidence values. They demonstrated the result on a simulated dataset, by taking example of four noisy inputs.

3 Object Detection

Simple background subtraction and Thresholding is ineffective in detecting the objects in various situations because of typical problems (as noted before) in both the domains. We employ a fusion based enhanced and efficient detection strategy using both visible and thermal imagery, which is well suited to handle typical problems in both the domains. Our approach is based on the use of gradient information along with background subtraction, as proposed and demonstrated in [5] but differs in the sense that we don't use computational intensive techniques for real timeliness. Additionally we take a fusion approach with visible spectrum video based on mutual agreement between the two modalities.

Since the algorithm requires registered imagery from the two sensors, we initialize the system by manually selecting four corresponding feature points from a pair of thermal and visible images. A homography matrix created from these points is used to register the thermal and visible images. First of all, localized regions of-interest (ROIs) are identified in both domains by applying standard gaussian background-subtraction, which generally produces regions that encompass the entire foreground object with surrounding halo in IR and shadows in visible, if present. The statistical background model for each pixel (in thermal or visible intensity) is created by computing *weighted* means and variances and the foreground pixels in the ROI then obtained using the squared Mahalanobis Distance by using following equation:

$$ROI(x,y) = \begin{cases} 1 & \dfrac{\left(I(x,y)-\mu(x,y)\right)^2}{\sigma(x,y)^2} \succ 100 \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

Now, at this step we examine the ROI's from both the domains, to get fused ROI that will be used to cue further processing in both the domains. Since the ROIs will include shadow in visible and halo in IR (if present) along with the foreground objects, taking intersection of both ROIs, will eliminate regions that are not present in both the modalities (like shadows, noise, etc.). However, if either of the sensor is performing poorly, either due to malfunctioning or environmental conditions, taking intersection will degrade the output of other sensor as well. Hence, we take the intersection only when both the ROIs have reasonable amount of mutual agreement in detecting foreground regions. Otherwise, we continue processing with the original ROIs separately for each domain and leave the detection of noise region at later stage. For determining the mutual agreement in both modalities, we use the following ratio (R), defined as:

$$R = \frac{P_{(1,1)}}{P_{(1,0)} + P_{(0,1)}} \tag{2}$$

where $P(x,y)$ is the total sum of pixels whose visible classification is x and whose infrared classification is y. Therefore, R is the ratio of the agreed foreground pixels to the total disagreed pixels. Now if R is greater than a predefined threshold, we assume that there is a high degree of mutual agreement in both the modalities, and we choose the fused ROI for further processing.

We again examine the difference image in each domain within resultant ROI in an attempt to extract gradient information corresponding only to the foreground object. Sobel operator is applied to calculate foreground gradient magnitudes from difference image and background gradient magnitudes. As proposed in [5] a gradient map is formed by taking pixel wise minimum of the normalized foreground gradient magnitudes and the normalized foreground-background gradient-difference magnitudes (as shown in equation 3), preserving the foreground object gradients that are both strong and significantly different from the background.

$$GradientMap = \min\left(\frac{\| \langle Ix, Iy \rangle \|}{\max}, \frac{\| \left(\langle Ix - BGx \rangle, \langle Iy - BGy \rangle \right) \|}{\max} \right) \tag{3}$$

By Thresholding the gradient map and applying morphological operations like closing and dilation, we obtain blobs corresponding to actual foreground objects (without halo or diffused shadows). The approach is equally applicable to both thermal and visible imagery. Figure 1 shows the output of the various steps of object segmentation applied to an infrared image having halo effect.

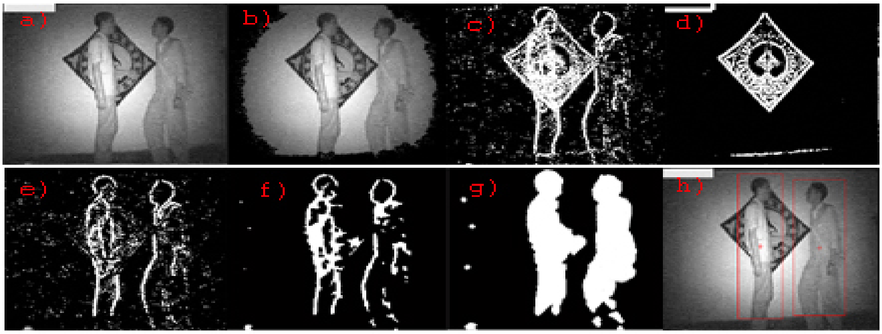

Fig. 1. Segmentation Output shown for infrared image with halo effect a) original image b) ROI c) Foreground gradient d) Background gradient e) Gradient map f) after Thresholding g) blobs h) objects detected

4 Target Tracking

Achieving better trajectory accuracy and continuity is of great importance for the successive steps of behavior understanding performed by a surveillance system. In particular, the trajectories of the objects in the scene have to be analyzed to detect suspicious events [11]. Tracking takes place at two levels. In the first level of the tracking procedure the system matches the blobs detected in the current frame with those extracted in the previous frame. Second level of tracking takes place at fusion level, where the objects (combination of one or more blobs) are tracked, using a fusion filter to obtain fused estimate of the object state. For getting an estimate of segmentation output and reliability, we compute certain measurement parameters for blobs, which are defined as follows:

1. "**Appearance Ratio (AR)**": Let D be the *difference map obtained as the absolute difference between the current frame and a reference image with T as threshold to binarize D , and let B_j be the j-th blob extracted from the sensor, then the Appearance Ratio for that blob is defined as*

$$AR(B_j) = \frac{\sum_{x,y \in B_j} D(x,y)}{|B_j| \times T} \tag{4}$$

where $|B_j|$ *is the number of pixels of the blob* B_j. The value of AR is proportional to the strength of the segmented blobs from each sensor. A low AR value indicates that the pixel intensity in the blob region has barely crossed the threshold. Thus AR value can be compared to determine which sensor is more informative

2. "**Overlapping**": *Overlapping O(a,b), between blobs a and b, is defined as:*

$$\begin{aligned} Omax(a,b) &= Maximum(\ IA(a,b)\,/\,A(a),\ IA(a,b)\,/\,A(b)) \\ Omin(a,b) &= Minimum(\ IA(a,b)\,/\,A(a),\ IA(a,b)\,/\,A(b)) \end{aligned} \tag{5}$$

where A(i) is the area of the i^{th} blob's bounding box, and *IA(a,b)* is the intersection area between them. These two factors are used in matching blobs.

3. "**Resemblance**": between two blobs is estimated with respect to the degree of match between two blobs (using *Omin*) and similarity factor. R(a,b) is defined as:

$$R(a,b) = Omin(a,b) \times [\ 1 - [Abs(Aa\text{-}Ab)\,/\,Maximum(Aa,\ Ab)] \tag{6}$$

4. "**Confidence (C)**": It gives the persistence of a blob over time and is defined by the following equation:

$$C(a) = (\sum_{b=0}^{n} R(a,b) \times C(b)) + 1 \tag{7}$$

where *a* is the new blob, *b is* the preceding blob, and *n* the number of preceding blobs that matched to the present one. As seen from the equation, the confidence on matching from *t-1* to *t* increases if the blob has been tracked for a long time and the resemblance from two time steps is large. Note that the minimum value of confidence of any blob is 1, which is in case of its first appearance in the scene.

Initially an object can be made up from an isolated blob or many closer blobs. The system first of all matches the current set of blobs with the objects detected in the previous frames by simple spatial occupancy overlap tests between the predicted locations of objects and the locations of blobs in current frame. The maximum overlapping factor (*Omax*) is used for this purpose. The system then establishes correspondence between the individual blobs (of current and previous frames) that correspond to same object. This is done by maintaining a list of blobs in the previous frames that correspond to each object. Subsequently, specific parameters like resemblance and confidence factor are calculated for each blob. The object's confidence is computed as the average of the confidence of individual blobs comprising the object. The appearance ratio for an object is calculated by summing up the numerator and denominator for each individual blobs and then dividing. The confidence C and appearance ratio AR is used in the fusion process to estimate the measurement accuracy of each sensor for extracted objects.

5 Fusion Process

We employ second order Kalman filter to model the motion of each object in the scene. The fusion procedure maintains its own list of targets. In the fusion process, the fused estimate should be more biased by accurate measurements and almost unaffected by inaccurate or malfunctioning ones. An intelligent fusion algorithm based on fuzzy logic techniques is designed to obtain fused measurement data (for each object). The main advantages derived from the use of fuzzy logic techniques with respect to traditional schemes are the simplicity of the approach, the capability of fuzzy systems to deal with imprecise information, and the possibility of including heuristic knowledge about the phenomenon under consideration [10].

The reliability of the sensors is estimated by two input parameters, the Appearance Ratio (AR) and Confidence (C). AR value reflects the strength of segmentation output from each sensor at current instance. The value of C also reflects on the temporal consistency of the sensor in maintaining good detection of a particular object. Also the confidence for an object detected as a single blob will be more than the object detected in fragemented parts (blobs).

Figure 2 shows the Hybrid Fuzzy logic-Kalman Fusion filter. A separate fuzzy inference system (FIS) is employed to monitor each channel and assigns suitable weights to each sensor's filtered estimate. Based on the values of the variables *C* and *AR,* the FIS assigns a weightage *w*, on the interval [0,1], to each of the KF's outputs. This value reflects the reliability of the sensor's measurement and it acts as a weight that tells the defuzzifactor, the confidence level at which it should take each KF's output value.

Each FIS was implemented using two inputs, the current value of *C* and *AR*; and one output, the weight *w*. For *C* and *AR*, we consider three fuzzy sets: ZE=zero, S=small, L=large. The membership function for C and AR are shown in figure 3. For

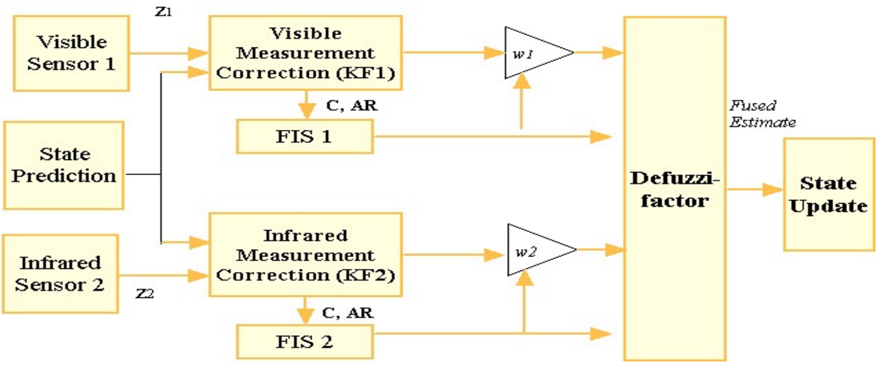

Fig. 2. A Hybrid Fuzzy logic-Kalman Fusion Filter

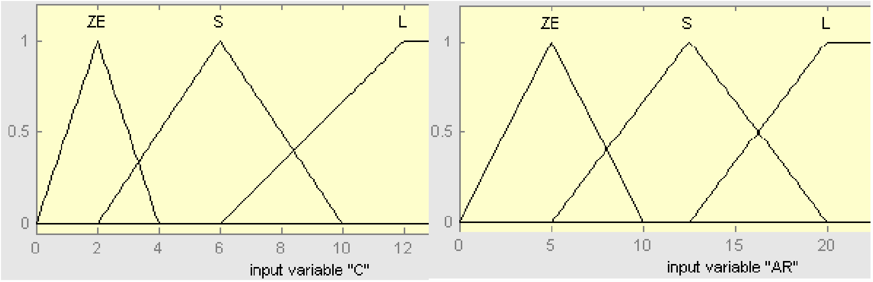

Fig. 3. Membership function for C and AR

the output *w*, three fuzzy singletons were defined with the labels: G=1=good, AV=0.5=average and P=0=poor. Thus the fuzzy rule base of each FIS comprises of following nine rules:

1. If *C*=ZE, and *AR*=ZE, then *w*=P
2. If *C*=ZE, and *AR*=S, then *w*=P
3. If *C*=S, and *AR*=ZE, then *w*=P
4. If *C*=ZE, and *AR*=L, then *w*=AV
5. If *C*=S, and *AR*=*S*, then *w*=AV
6. If *C*=L, and *AR*=ZE, then *w*=*AV*
7. If *C*=S, and *AR*=*L*, then *w*=G
8. If *C*=L, and *AR*=*S*, then *w*=G
9. If *C*=L, and *AR*=*L*, then *w*=G

The above rules are based on two simple heuristic considerations. First, if both C and AR are large for an extracted object from a sensor, it implies that the sensor's filtered estimate is highly reliable. Second, if both of these values are near to minimum, the output is unreliable. Thus, using the compositional rule of inference sum-prod, the FIS calculates the weight, which tells the defuzzifactor at what confidence level it should take each output. Note that this method of fusion is suitable for any number of sensors.

6 Experimental Results

For experiments, we used the Sony TRV65 Hi8 Camcorder with 37mm 1000nm IR filter that allows recording of daytime video (visible spectrum) and nighttime images in indoor situation. For outdoor situation we used MATIS thermal camera with InSb detector (320x284) and spectral range of 3-5 µm, for capturing good quality infrared images. The program implementation was done in Matlab 7.0. We tested our approach at every stage to analyze the improvement in the performance obtained by combining visible and infrared imagery.

Object Detection: The object detection approach is robustly able to detect objects across a wider range of environmental conditions than is possible with standard approaches as demonstrated in [5]. Here we have also tested the method over different ranges of IR sensors (with varying degree of noise, halo effect etc) suitable for indoor and outdoor surveillance. We collected samples of IR images from three sensors ranging from high, medium and low quality. Figure 4 shows segmentation result on sample images from these sensors taken in indoor and outdoor situations.

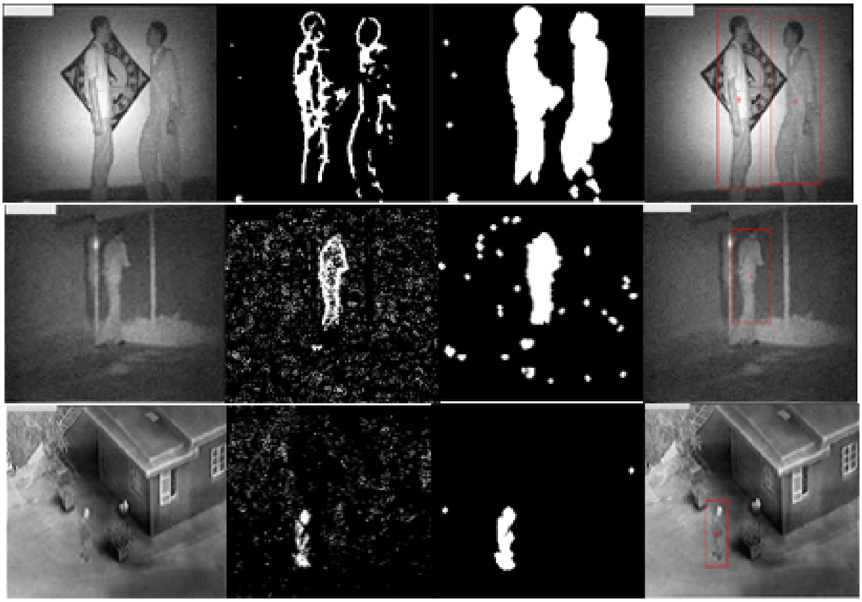

Fig. 4. First column shows the original IR images, second column shows the corresponding gradient map, third column shows the blobs extracted and the fourth one show the objects detected in the images

The IR images in the first and second row are night shot images (from Sony Camcorder) taken in indoor situations. The IR image in first row contains Halo around people and the second row image is extremely noisy. The third row shows an outdoor situation where the person’s body is quite insulated by clothing and only the head portion appears as hot spot. In spite of these challenges, the output shows properly

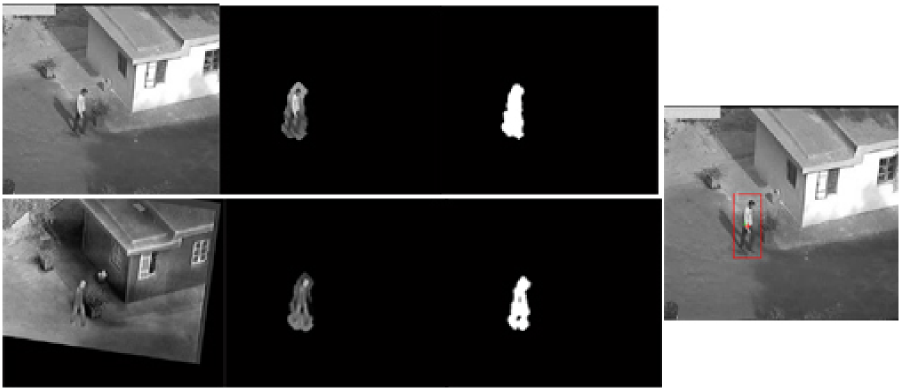

Fig. 5. First row shows the image in visible domain having shadow and the second row shows the corresponding IR image, after registration. The second column shows the Fused ROI in both domain and the third column shows the blob extracted. Finally the detected object region (shown in visible image) does not contain shadow.

segmented out objects. In Figure 5, an outdoor situation of one pedestrian walking near a building is presented and it shows that the IR image can be helpful in removing shadows from the visible image.

Object Tracking: For comparison of accuracy in tracking the trajectory, the following performance measures were adopted:

$$J_{zv} = \sqrt{\frac{1}{n}\sum_{k=1}^{n}\left(za_k - z_k\right)^2} \tag{8}$$

$$J_{ze} = \sqrt{\frac{1}{n}\sum_{k=1}^{n}\left(za_k - \hat{z}_k\right)^2} \tag{9}$$

Where za_k is the actual value of the position; z_k is the measured position; and $\hat{z}_k$ is the estimated position at an instant of time k. Figure 6 (appendix) shows a pedestrian being tracked with fused measurement of centroid position (shown with red cross) and fused estimate (with green cross) is shown in visible imagery. Actual position was calculated by manually segmenting the pedestrian. Table 1 shows the performance measures obtained by using only visible, only thermal and using both modalities.

Table 1. Comparison of tracking accuracy obtained by using only visible, only thermal and using both

Sensor	J_{zv}	J_{ze}
Only Visible	2.04	2.92
Only Infrared	3.35	3.20
Visible and Infrared (fused)	1.87	2.45

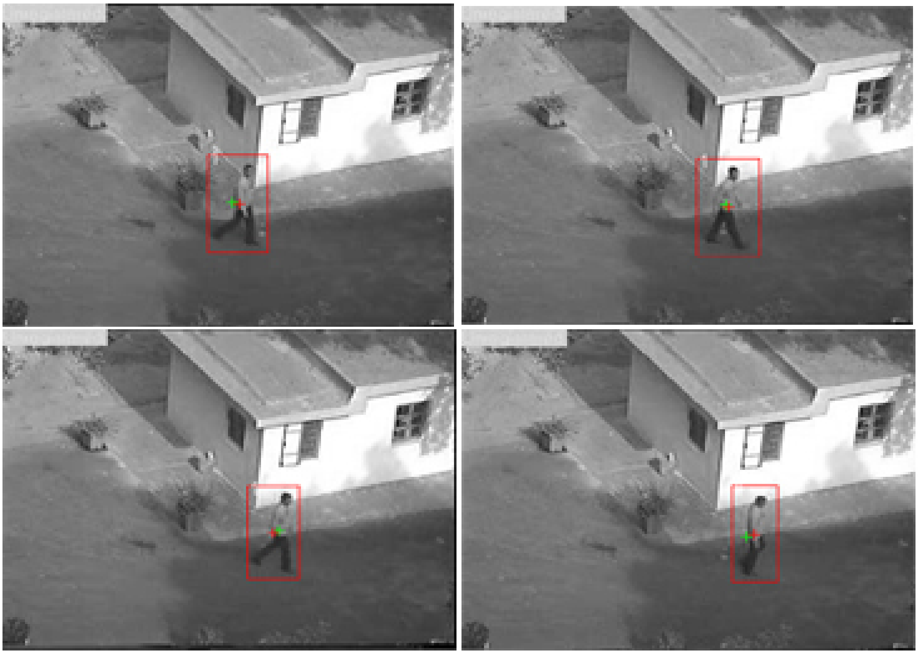

Fig. 6. A pedestrian being tracked in outdoor situation

Since it is a daytime situation with sufficient lighting and temperature difference in the environment, the two sensors are reporting a track similar to the ground truth. Nonetheless, a better result is obtained through data fusion. We haven't shown obvious case of night situation, where visible sensor fails completely and the fused output is according to the infrared sensor only.

7 Conclusion

In this paper, we presented the framework for combining visible and IR for robust object detection and accurate tracking in a surveillance system. The problems that arise in each domain and the potential of combining both modalities in addressing these problems were discussed. Fusion approach for combining information from visible and infrared source at segmentation level and tracking level was discussed in detail. An enhanced object detection strategy is implemented with efficient fusion based approach to handle typical problems of both the domains The following are the contributions of our work. The novelty of our work lies in using Fuzzy logic based-Kalman filtering technique to track objects and obtain fused estimate according to the reliability of the sensors. Suitable measurement parameters are identified to automatically estimate the measurement accuracy of each sensor so that they can be fused in a weighted manner.

References

1. Conaire, C., O. Transfer report-Phd register: Thermal Infrared and Visible Spectrum Fusion for Multi-modal Video Analysis. Dublin City University. (July 28, 2005)
2. Hu, W., Tan, T., Wang, L., and Maybank., S. A survey on visual surveillance of object motion and behaviors. IEEE Transactions on Systems, Man and Cybernetics, 34(3):334-350, August (2004)
3. Cucchiara., R. Multimedia surveillance systems. In VSSN '05: Proceedings of the third ACM international workshop on Video surveillance & sensor networks, New York, NY, USA, pages 3-10 (2005)
4. Bhanu., B., and Han., J. Kinematic-based human motion analysis in infrared sequences. In Proceedings Workshop Applications of Computer Vision, pages 208–212 (2002)
5. Davis, J., and Sharma, V. Robust detection of people in thermal imagery. In Proceedings of International Conference on Pattern Recognition., pages 713–716 (2004)
6. McDaniel, R., Scribner, D., Krebs, W., Warren, P., Ockman, N., McCarley, J. Image fusion for tactical applications. Proceedings of the SPIE - Infrared Technology and Applications XXIV, 3436, 685-695(1998)
7. Torresan, H., Turgeon, B., Ibarra-Castanedo, C., Hébert, P., and Maldague, X. Advanced Surveillance Systems: Combining Video and Thermal Imagery for Pedestrian Detection. In Proceedings of SPIE, Thermosense XXVI, volume 5405 of SPIE, pages 506–515, (2004)
8. Davis, J., and Sharma, V. Fusion-Based Background-Subtraction using Contour Saliency. Computer Vision and Pattern Recognition, 20-26 (June, 2005)
9. Snidaro, L., Niu, R., Varshney, P.K., and Foresti, G.L. Automatic camera selection and fusion for outdoor surveillance under changing weather conditions. IEEE Conference on Advanced Video and Signal based Surveillance, Florida, pp. 364–370 (2003)
10. Escamilla-Ambrosio, P.J., Mort, N. A Hybrid Kalman Filter - Fuzzy Logic Architecture for Multisensor Data Fusion. Proceedings of the 2001 IEEE International Symposium on Intelligent Control , pp. 364-369 (2001)
11. Regazzoni, C., Ramesh, V., and Foresti, G.L. Special issue on video communications, processing, and understanding for third generation surveillance systems. Proceedings of the IEEE, 89(10), 2001.