Feature-Level based Video Fusion for Object Detection

Anjali Malviya[1], S. G. Bhirud[2]

**[1] Dept. of IT, TSEC, Mumbai University
Mumbai, Maharashtra, India**

**[2] Dept. of Computer Engg., VJTI
Mumbai, Maharashtra, India**

Abstract

Fusion of three-dimensional data from multiple sensors gained momentum, especially in applications pertaining to surveillance, when promising results were obtained in moving object detection. Several approaches to video fusion of visual and infrared data have been proposed in recent literature. They mainly comprise of pixel based methodologies. Surveillance is a major application of video fusion and night-time object detection is one of most important issues in automatic video surveillance. In this paper we analyse the suitability of a feature-level based video fusion technique that overcomes the drawback of pixel-based fusion techniques for object detection.

Keywords: video fusion; feature-level-fusion.

1. Introduction

Multisensor fusion attempts to combine the information from all available sensors into a unified representation. In other words, it refers to any stage in the integration process where there is an actual combination (or fusion) of different sources of sensory information into one representation. Some of the advantages to multisensory fusion are improved detection, increased accuracy, reduced ambiguity, robust operation, and extended coverage. To illustrate how these advantages come about, relationship among sensors are categorized into three types of relations, complementary, competitive, and cooperative. Moreover, fusion can take place at pixel, feature or decision level. There has been an explosion of applications in multisensor fusion and integration. Multiple-sensor based visual surveillance systems can be extremely helpful because the surveillance area is expanded. Tracking with a single sensor easily generates ambiguity due to limitations of object capturing, especially with insufficient light. This ambiguity may be eliminated from another view via other sensor.

The importance of video surveillance techniques [1-2] has increased considerably since the latest terrorist incidents. Safety and security have become critical in many public areas, and there is a specific need to enable human operators to remotely monitor activity across large environments such as transport systems (railway transportation, airports, urban and motorway road networks, and maritime transportation), banks, shopping malls, car parks, and public buildings, industrial environments, and government establishments (military bases, prisons, strategic infrastructures, radar centers, and hospitals). Modern video-based surveillance systems [2] employ real-time image analysis techniques for efficient image transmission, color image analysis, event-based attention focusing, and model-based sequence understanding. Moreover, cheaper and faster computing hardware combined with efficient and versatile sensors create complex system architectures; this is a contributing factor to the increasingly widespread deployment of multi-camera systems. These multi-camera systems can provide surveillance coverage across a wide area, ensuring object visibility over a large range of depths. They can also be employed to disambiguate occlusions. Techniques that address handover between cameras (in configurations with shared or disjoint views) are therefore becoming increasingly more important. Events of interest (identified as moving objects and people) must be then coordinated in the multi-view system, and events deemed of special interest must be tracked throughout the scene. Wherever possible, tracked events should be classified and their dynamics (sometimes called behavior) analyzed to alert an operator or authority of a potential danger.

In the development of advanced visual-based surveillance systems, a number of key issues critical to successful operation must be addressed. The necessity of working with complex scenes characterized by high variability requires the use of specific and sophisticated algorithms for video acquisition, camera calibration, noise filtering, and motion detection that are able to learn and adapt to changing scene, lighting, and weather conditions. Working with scenes characterized by poor structure requires the use of robust pattern recognition and statistical methods. The use of clusters of fixed cameras, usually grouped in areas of interest but also scattered across the entire scene, requires automatic methods of compensating for chromatic range differences, synchronization of acquired data (for overlapping and non overlapping views), estimation of correspondences between and among overlapping views, and registration with local Cartesian reference frames.

However, visual surveillance using multi cameras also brings problems such as camera installation, camera calibration, object matching, automated camera switching, and data fusion.

The image fusion techniques implemented earlier comprised of essentially pixel-level fusion. For video fusion, we explore feature-level fusion methodologies, along with pixel-level-fusion. Fusion at the feature level requires extraction of objects (features) from the input images. These features are then combined with the similar features present in the other input images through a pre-determined selection process to form the final fused image. Since one of the essential goals of fusion is to preserve the image features, feature level methods have the ability to yield subjectively better fused images than pixel based techniques.

2. Feature Level Fusion

Image fusion algorithms can be categorized into low, mid, and high levels [3]. In some literature, this is referred to as pixel, feature, and decision levels. Methods using pixel level either use arithmetic operations (like addition, subtraction) on corresponding pixel intensity from different input images or use the frequency domain. Using the frequency domain, the input images are first transformed in the frequency domain using various pyramid based methods like Laplacian, or Wavelet transforms. After transformation, algebraic operations are performed on the input images fusing them to one image. Then, that image is inverse transformed to the final fused image.
Feature level methods are the next stage of processing where image fusion may take place. Fusion at the feature level requires extraction of objects (features) from the input images. These features are then combined with the similar features present in the other input images through a pre-determined selection process to form the final fused image. Since, one of the essential goals of fusion is to preserve the image features, feature level methods have the ability to yield subjectively better fused images than pixel based techniques. A schematic of feature level fusion is shown in Figure 1. The typical algorithms used are feature-based template methods (like edge enhancement), Artificial Neural Networks, and knowledge based approaches.

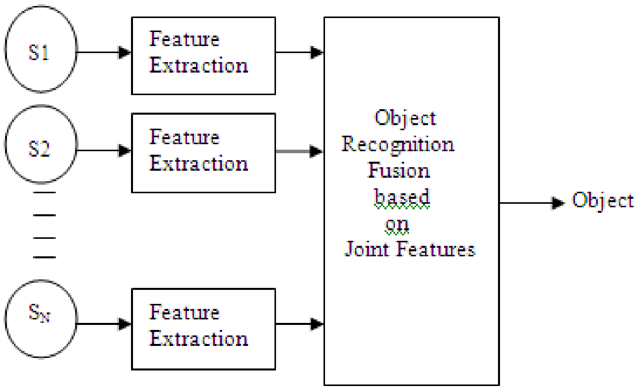

Fig. 1 Schematic of Feature Level Fusion

3. Object Detection using Feature-Level Video Fusion

Night time vision is a primary need of video fusion for surveillance application. The literature survey on video fusion drew our attention towards the problem of noise in IR video, which proves to be a huge deterrent in obtaining high quality fused videos, thereby affecting the surveillance application. We aimed at developing a procedure that could address IR video de-noising and at the same time help in pedestrian detection problem in night-time environment. Unlike most of the work on IR video de-noising, this method does not require static background assumption and Gaussian noise assumption. It involves three steps, IR video de-noising, object detection (pedestrian in this case) in IR video and visual-infrared video fusion.

4. Methodologies

Additive and multiplicative noise is an unwanted component of videos. They can occur as Gaussian noise or film grain noise and may have undesirable effects on surveillance applications. The first stage of most video processing techniques is noise removal but mere usage of spatial noise removal techniques can only give limited filtering performance [4]. Improved performance can be achieved by considering a sequence of previous and/or subsequent image frames for filtering, leading to a spatio-temporal filtering. We use a 3D window around pixel (x, y, t) for our filtering, as proposed in [5]. If noise in preceding and current frame is additive white Gaussian noise (AWGN), taking linear average of the pixels in 3-D window gives good results for de-noising.
Pedestrian motion effects should be taken into account when filtering in order to reduce temporal filtering artifacts such as blurring. Image regions that include pedestrian motion in preceding frames and /or current

frames should not be taken into account while filtering. For this purpose, we use brightness (or intensity value) threshold T1 to determine pixels which potentially belong to the pedestrian, because we assume that the pedestrian has higher temperature than the environment, and hence the pixels corresponding to the pedestrian in infrared frames would be brighter than the background environment.

The object (pedestrian in this case) is segmented from the background, using the thermal-image features of pedestrian. As pedestrian region is brighter than the background in infrared video, the regions can be segmented according to their brightness. Shape recognition algorithms are then used to separate the pedestrian regions from other false detections. The segmentation algorithm proposed by Adams/Bischof [6] is used. The image is first searched for a “seed” pixel belonging to “pedestrian-type,” in every frame. If found one, then we segment one candidate-region of pedestrian by applying the seeded region growing algorithm proposed in [6] and continue such search until all candidate-regions are segmented.

The brightness-only information is not robust in detection of pedestrian, as there are other thermal emitters in the environment which can lead to false detection. In order to improve the robustness of detection, we fuse color information and shape information. We use area feature of bright region to detect pedestrian. If area of the bright region is greater than a threshold value and height/width ratio is in the previously-established range, then the region is regarded as a pedestrian else it is regarded as noise region.

The bright regions in infrared frames correspond to detected pedestrian. At the end of our detection method, infrared frames are fused with visual frames to provide visual context. We fuse the frames by adding an increment T_6 to pixel’s RGB values in visual frames at the corresponding geometric position, by which pedestrian region in visual video can be made more visible to human vision.

5. Implementations and Results

The dataset used is “AIC Thermal/Visible Night-time Dataset” which contains two video sequences, one in the visible spectrum and one in thermal infrared (Input 1). Both are compressed into AVI format and contain 527 frames each. It was captured from a balcony in Dublin City University campus, Ireland [7]. Figure 2 shows four frames from the IR video and the corresponding four frames from the visible video are shown in Figure 3. Figure 4 & 5 show the classification of noisy and pedestrian pixels, in the IR frame, and the denoised IR frames. Segmenting candidate region using seeded region growing can be seen in Figure 6. Figure 7and Figure 8 show the final fused image.

The methodology, is also tested on a day-time video (Input 2) obtained through visible camera and IR camera respectively.

6. Conclusion

Pedestrian image region is brighter than background in infrared video, thus the regions can be segmented according to their brightness [Input 1]. This is however not robust in detection of pedestrian, as there are other thermal-emitters in environment which can lead to false detection [Input 2]. Therefore to improve robustness of detection, we fuse color information (brightness) and shape information. In the fused tracking video frames [Fig. 7 and 8], detected pedestrian have been marked by a red rectangle, after enhancing its brightness. As a result this method provides a more visualized pedestrian detection result for human vision. However the algorithm fails to detect the object distinctly in the day time video where there are other bright objects in the IR video [Fig. 10 – Fig. 14].

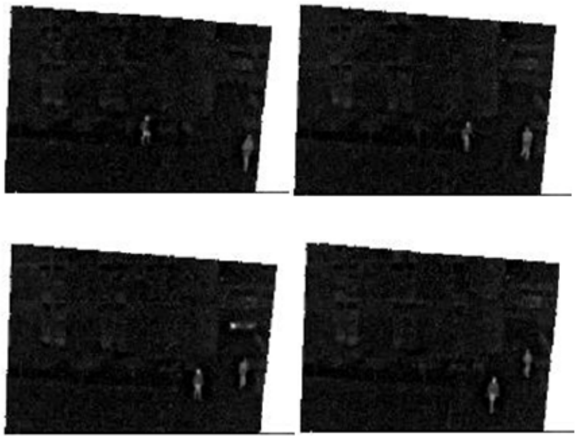

Fig.2 IR Video Frames from Input 1

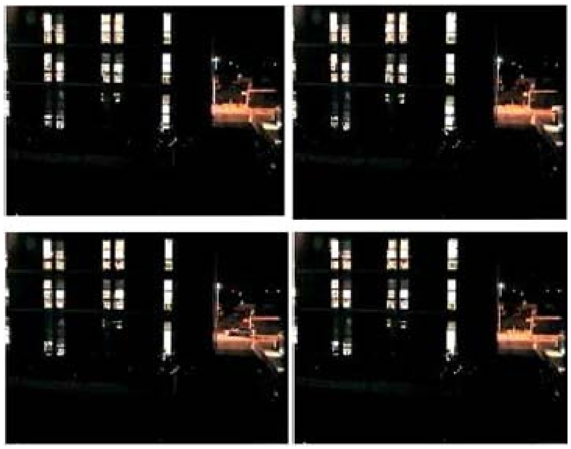

Fig.3 Visible Video Frames from Input 1

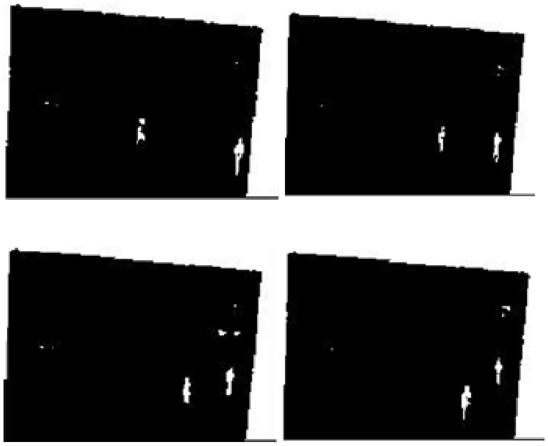

Fig. 4 Classification of IR Frames into Pedestrian and Noisy Pixels

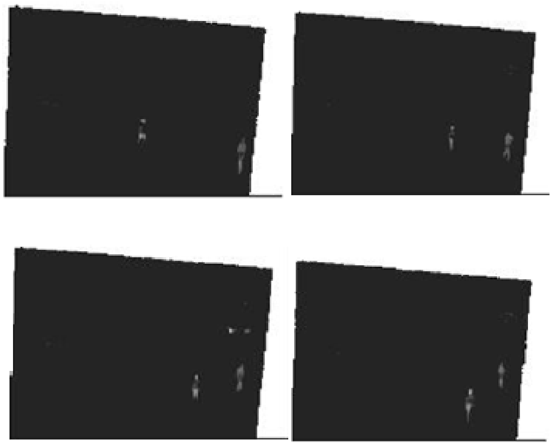

Fig.5 De-noising of IR Frames

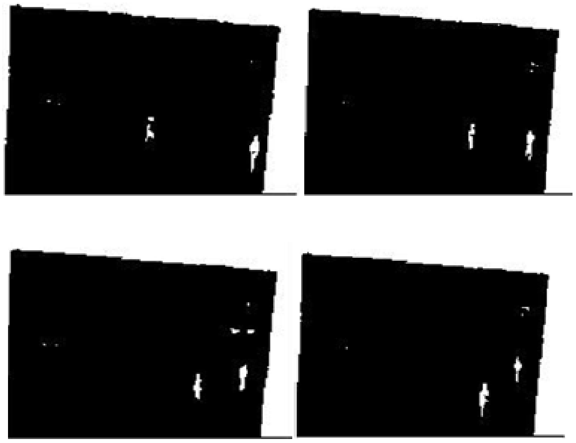

Fig.6 Seeded Region Growing

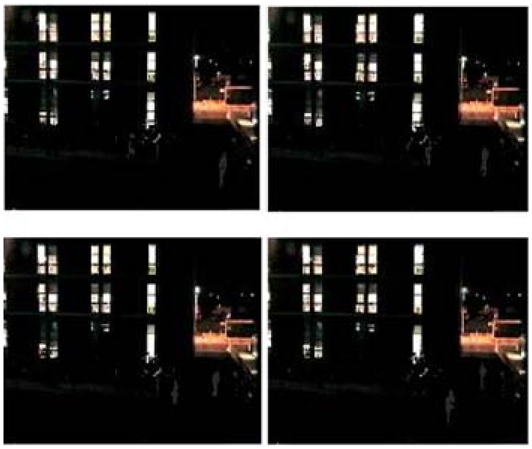

Fig. 7 Fused frames with the Pedestrian Regions

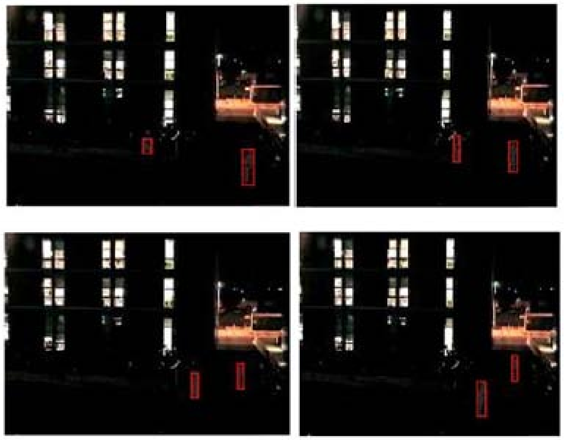

Fig. 8 Frames from the Fused Video

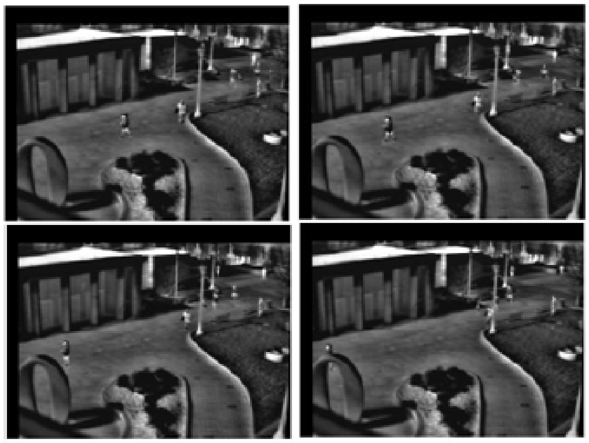

Fig.9 IR Video Frames from Input 2

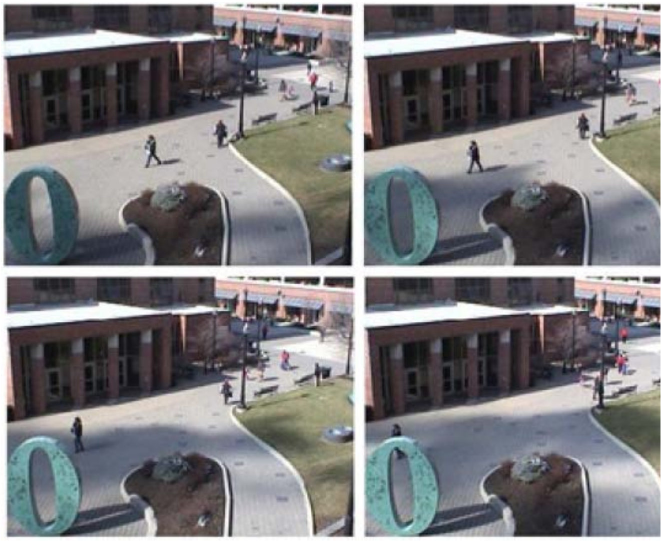

Fig. 10   Visible Video Frames from Input 2

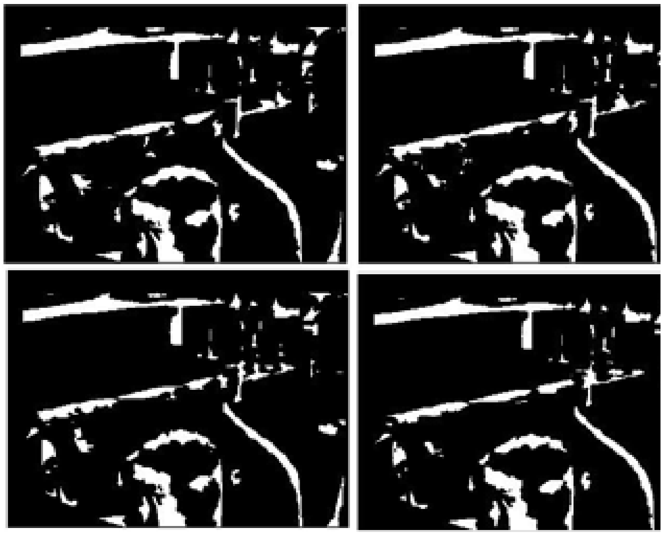

Fig. 11   Classification of IR Frames into Pedestrian and Noisy Pixels

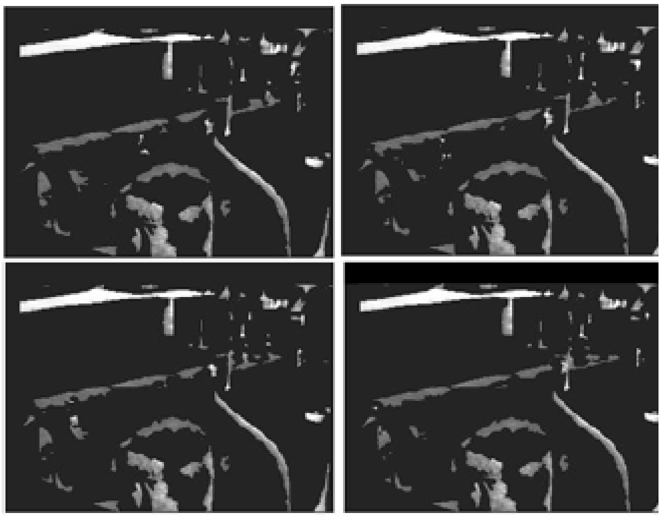

Fig.12   De-noising of IR frames

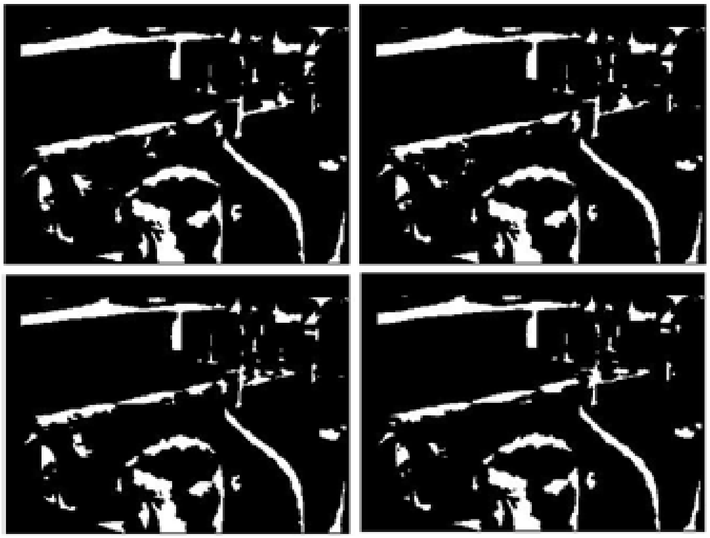

Fig.13   Seeded region growing

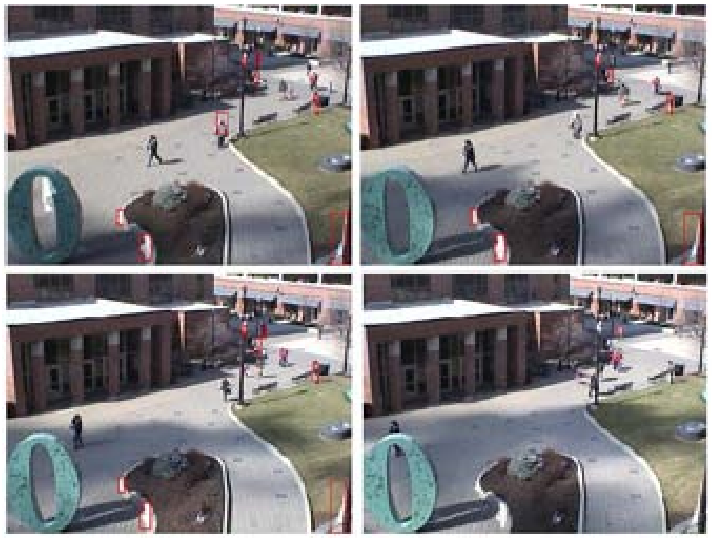

Fig. 14 Fused Image

References

[1] H.M. Chen, S. Lee, R.M. Rao, M.A. Slaman, and P.K. Varshney, “Imaging for concealed weapon detection,’’ IEEE Signal Processing Mag., vol. 22, no. 2, pp. 52–61, Mar. 2005.

[2] G.L. Foresti, C.S. Regazzoni, and R. Visvanathan, “Scanning the issue/technology - Special issue on video communications, processing and understanding for third generation surveillance systems,’’ Proc. IEEE, vol. 89, no. 10, pp. 1355–1367, Oct. 2001.

[3] Yaonan Wang, Multisensor Image Fusion: Concept, Method and Applications, Faculty of Electrical and Information Engineering, Hunan University, Changsha, 410082, China.

[4] Brailean J., Kleihorst R., Efstratiadis S., Katsaggalegos A., and Lagendijk A. Noise Reduction Filters for Dynamic Image Sequences: A Review, Pcoc. Of IEEE 83 (9), (1995) 1272-1292

[5] M. Kemal G¨ull¨u, O˘guzhan Urhan, and Sarp Ert¨urk, Pixel Domain Spatiotemporal Denoising for Archive Videos. ISCIS 2006, LNCS 4263, pp. 493-502, 2006.

[6] R. Adams and L. Bischof, Seeded region growing, IEEE Trans. Pattern Anal. Machine Intell., vol. 16, no. 6, pp. 641C647, 1994

[7] O Conaire, C., O Connor, N.E., Cooke, E., Smeaton, A.F.: Comparison of fusion methods for thermo-visual surveillance tracking. International Conference on Information Fusion (Fusion 2006)