Pattern Anal Applic (2017) 20:907–926
DOI 10.1007/s10044-017-0621-z

SURVEY

Fusion of thermal infrared and visible spectra for robust moving object detection

Emna Fendri[1] · Rania Rebai Boukhriss[2] · Mohamed Hammami[1]

Received: 10 October 2016 / Accepted: 8 April 2017 / Published online: 18 April 2017

Abstract The detection of moving objects is a crucial step for many video surveillance applications whether using a visible camera (VIS) or an infrared (IR) one. In order to profit from both types, several fusion methods were proposed in the literature: low-level fusion, medium-level fusion and high-level fusion. The first one is the most used for moving objects' detection in IR and VIS spectra. In this paper, we present an overview of the different moving object detection methods in IR and VIS spectra and a state of the art of the low-level fusion techniques. Moreover, we propose a new method for moving object detection using low-level fusion of IR and VIS spectra. In order to evaluate quantitatively and qualitatively our proposed method, three series of experiments were carried out using two well-known datasets namely "OSU Color-Thermal Database" and "INO-Database"; the results of these evaluations show promising results and demonstrate the effectiveness of the proposed method.

Keywords Thermal infrared spectrum · Visible spectrum · Moving object detection · Image fusion

✉ Rania Rebai Boukhriss
rania.rebai@hotmail.fr

Emna Fendri
fendri.msf@gnet.tn

Mohamed Hammami
mohamed.hammami@fss.rnu.tn

[1] MIRACL-FS, Sfax University, Road Sokra Km 3 BP 802, 3018 Sfax, Tunisia

[2] MIRACL-ISIMS, Sfax University, Sakiet Ezzit, Sfax, Tunisia

1 Introduction

During the last few decades, computer vision has acquired a growing interest due to the need for security, and offered several software applications for video surveillance in public and private sites, safety and traffic control and robotics, etc. Despite their diversity, these applications are all based on an essential step that is the detection of moving objects in the video stream [1]. In fact, the moving areas in a video stream often correspond to events on which a vision system must focus.

In the vision systems, we can use either a single type of camera (visible or infrared) or both. The reason behind the use of visible spectrum is the rich content offered by visible images, containing texture, strong edges, color and other information with low noise. However, using only visible sensor, the detection of moving objects can be limited by levels of darkness and luminosity, shadows, light reflections, camouflage and weather conditions such as fog, smoke, rain, snow, etc. In order to overcome these limitations and to correctly carry out surveillance in outdoor scenarios, many works propose to use IR sensor. In fact, an infrared camera captures the temperature emitted by objects. So it is nearly invariant to changes in ambient illumination [2] and provides valuable information at night and/or in poor visibility conditions. This characteristic enables it to be of great benefit to monitoring and surveillance systems, as it can operate on a 24-h basis, and it is reliable at detecting hot objects, such as people and vehicles, which are normally the primary objects of interest in surveillance [3]. However, such an infrared-based system may find it difficult to handle some information in certain situations. For example, during a hot sunny day, it will highlight almost the entire image, so it will provide a lot of hot areas or objects (in this case even the pavement will be seen as it emits heat). For this reason, the use of a VIS sensor with an IR sensor makes the vision systems more robust and enables them to function under varying lighting and climatic conditions, both day and night, in summer as well as in winter. Therefore, we could conclude that neither of these two sensors would perform very well alone in all situations. A fusion of IR and VIS spectra is interesting to solve more intricate situations. Since our aim is to perform correct moving object detection all over the day (morning, afternoon and night) for particular hot objects such as people and vehicles, we chose to use both the VIS and IR spectra.

The remainder of this paper is organized as follows: Section 2 provides the literature survey on moving object detection in IR and VIS spectra. Section 3 provides an overview of the literature related to IR–VIS fusion techniques. Section 4 describes the proposed method for moving object detection using the fusion of IR and VIS spectra. Quantitative and qualitative evaluations of our work are outlined in Sect. 5. Finally, our conclusion and future works directions are introduced in Sect. 6.

2 Literature survey of moving object detection in IR and VIS spectra

The detection of moving objects is a binary detection as it is performed to decide which parts of the frames (pixels or regions) belonging to moving objects to detect. The diversity of research is related to the complexity of the observed scene that presents a variety of challenges such as sudden and progressive illumination changes, shadows projected by the moving objects, ghost, camouflage and occlusion, etc. A great number of methods of moving object detection in IR and VIS spectra have already been proposed. The contributions reported in the literature can be classified in three main approaches according to the nature of the treatment they achieve: local approach, global approach and hybrid approach.

2.1 Local approach

In this approach, there are two categories of methods: motion-based methods and features-based methods.

2.1.1 Motion-based methods

The methods based on motion include several methods founded on inter-frame processing that are based either on inter-frames differences, background modeling or on the optical flow [4–6].

Inter-frame difference-based methods The inter-frame difference methods are used with videos captured either in the VIS spectrum [7–9] or in the IR spectrum [10, 11]. The basic method introduced by [12] is to calculate the absolute value of the difference between the intensities of the pixels (in gray) of each two successive frames; then, this difference will be compared to a decision threshold to obtain a binary mask of pixels in motion. The development efforts in this category of methods tend to improve not only the method of calculating the binarization threshold but also the manner of fixing the spacing between the successive frames [13]. The advantages of these methods are their high processing speed due to their low computational complexity for the extraction of moving pixels and their robustness in adapting with dynamic environments. However, they are sensitive to noise and sudden illumination changes. Furthermore, these methods fail to detect all pixels of the same moving object, in particular those inside the object.

Background modeling-based methods This category of methods is the most popular for moving object detection either in VIS [14–18] or in IR [19–21] spectra. The detection of a moving object is performed in two phases. The first phase builds a background model, where, the majority of the work uses the frames acquired offline and not containing any foreground objects. The second phase detects the moving pixels through the subtraction between the current image and the background model. However, the effectiveness of these methods depends on the background frame modeling and updating which is not easy to reach with complex scenes. Since the illumination conditions generally change faster than temperature, the background adaptation is generally less crucial in IR than in VIS spectra. Thus, the detection of a moving object with background modeling is easier in IR than in VIS because the IR spectrum is less sensitive to changes of lighting that of temperature [22].

Optical flow-based methods The optical flow is used for moving objects detection in a VIS spectrum [23, 24] and in an IR spectrum [11, 25]. This motion estimation method is achieved by translation vectors. Almost all of the calculation methods of optical flow are based on the assumption that a pixel of a sequence frame maintains constant brightness intensity during its movement. These methods work very well in changing environments and can detect moving objects with camera motion. However, they suffer from a high temporal complexity which makes it difficult to achieve a real-time detection without a specialized hardware. Therefore, most of the research is focused on the performance of the algorithm, ignoring the feasibility in practice [25].

2.1.2 Feature-based methods

These methods are widely used for the detection of objects of a known class, especially the pedestrian class, either in

IR or in VIS spectra. With the IR imagery, the detection of target objects is based on different features such as pixels intensity where the bright pixels of hot objects like pedestrians, animals or vehicles are selected by thresholding [26–28]. Furthermore, some of them use the symmetry of the contour [28, 29], size or aspect ratio (ratio of height and width) of the selected bounding boxes [29–31] as well as the histograms of oriented gradients [32–34]. These methods, however, do not work very well with the various challenges present in the IR spectrum such as background noise (hot spot) or change of temperature. Besides, the thresholds used to detect bright pixels are specific to the scene. In a VIS spectrum, many authors combine a set of features to detect a specific object in a monocular image. Wang et al. [35] proposed a human detection approach capable of handling partial occlusion using the Histograms of Oriented Gradients (HOG) combined with the Local Binary Pattern (LBP) as a feature set. These descriptors are learned from the training data using the linear SVM. However, this method cannot handle the articulated deformation of people. Likewise, Schwartz et al. [36] have presented a human detection method using edge-based features with texture and color information which are learned with the SVM classifier.

In an application of moving objects detection, there are various target objects such as animals, pedestrians and the different types of vehicles. Therefore, the feature-based methods are generally correlated with other motion-based methods [3, 31] or model-based (head, legs or car model) [29, 37] to improve the results of moving object detection and can handle with several object classes.

2.2 Global approach

The global approach methods are used in the IR [37–39] and VIS [40–42] spectra for detecting the target objects such as pedestrians or cars based on models built for the purpose. The used models are head, body, arms and legs of pedestrians or car models. The models of the human body (head, legs or arms) are mainly used in IR images for detecting pedestrians [29, 34, 37, 38]. Indeed, the detection of heads, by looking for bright regions having a circular shape in the IR images, a pedestrian present in a crowd or in occlusion can be detected separately [43]. Using only the VIS spectrum, Lin et al.[42] proposed a learning-based, sliding window-style approach for the problem of detecting humans by simultaneously segmenting human shapes and poses, and extracting articulation-insensitive features. The shapes and poses are segmented by an efficient, probabilistic hierarchical part-template matching algorithm, and the features are collected in the context of poses by tracing around the estimated shape boundaries. Likewise, Leibe et al. [41] proposed a method that combines the segmentation and the recognition into one process using an implicit shape model. In fact, the implicit shape model is formulated in a probabilistic framework allowing us to obtain a category-specific segmentation of objects such as cars. This segmentation can then in turn be used to improve the recognition results.

Although the results of the global approach methods are promising, many of them are focused on only one category of object (vehicle or pedestrian) due to many challenges such as the highly articulated body postures, viewpoint changes, varying illumination conditions and background clutter. Moreover, the correlation of these models requires a significant computing time.

2.3 Hybrid approach

The above-mentioned approaches have advantages and disadvantages. To develop robust algorithms for a moving object detection, the idea of the hybrid methods [44–49] is to use a combination of two or three methods of these approaches to benefit from the advantages of each one. Based on the IR imagery, the authors of [47] have proposed a hybrid multiresolution methodology for moving object detection by combining the background subtraction, inter-frame difference and the optical flow methods at different resolutions. Moreover, Yin et al. [48] have also proposed an infrared moving object detection method based on background modeling and inter-frame difference methods. In fact, they make an AND operation on the binary foreground results of the two methods to obtain the final moving regions. In [45, 46], the authors proposed foreground detection methods in a VIS spectrum by combining the background subtraction with the inter-frame differencing. In fact, they incorporate the inter-frame differences' methods in the background modeling step. As for Lu et al. [49], they used the three motion-based methods (inter-frame difference, background modeling and optical flow) sequentially in a VIS spectrum. In fact, they begin with the difference between successive frames to detect the foreground areas to calculate the optical flow. Then, they exploit the optical flow results to update the two modeled backgrounds: a long-run background and a short-run one. Most of the works using both IR and VIS spectra for the detection of moving objects are hybrid methods used for a better exploitation of the features of the two cameras. Indeed, the proposed methods in [3] and [31] are based on motion and features for a better detection of moving pedestrians in IR and VIS images. In fact, they combined these two methods sequentially. Firstly, they used background modeling for the detection of moving regions. Secondly, they selected the regions corresponding to pedestrians relying on some features such as bright pixels, size or aspect ratio (ratio of height and width) and histograms of the selected bounding boxes.

2.4 Discussion

In this section, we reviewed the different methods proposed for the detection of moving objects in IR and VIS spectra. We classified these methods into three main approaches based on the nature of the treatment they achieved: the local, global and hybrid approaches. The objective is to have an effective method for the moving object detection in terms of temporal and spatial complexity, fast in terms of adaptability and independent of the moving objects speeds and sizes. Furthermore, as our aim was to improve the results of a moving object detection all over the day and during varying lighting and climatic conditions, we decided to use the VIS and IR spectra simultaneously. The study of the state of the art of moving object detection methods shows that the background modeling method is well suited to our goals. The originality of our proposed method is in fusing the IR and VIS spectra for moving object detection by applying background modeling-based method, incorporating the principle of the inter-frame differences' methods in the background modeling stage. In the following section, the state of art of IR–VIS fusion methods was presented.

3 State of art on fusion methods of VIS and IR spectra

Sensor fusion is a research area addressed by several authors. Most of them have limited their definition to the context of their application domain. During the recent years, the sensor fusion has become an increasingly important research field in computer vision systems. In fact, a fusion of the information provided by VIS and IR cameras for the task of moving object detection would solve difficult complementary situations, which any system based only on one type of camera could not solve on its own [50, 51]. In many multisensor systems, fusion algorithms should significantly reduce the amount of raw data that need to be presented or processed without loss of information content and provide an effective way of information integration [2], as well.

3.1 Fusion levels

Over the years, numerous techniques have been developed to address the growing need for sensor fusion. These techniques can be classified on the basis of the processing level where the fusion takes place [2, 22, 50]. There are three main levels where image fusion may take place and they include:

Low-level fusion In the low-level fusion, also called signal, data or pixel-level fusion, raw images obtained from the sensors are combined to produce a new fused image before applying any information extraction algorithm. In our context, the IR and VIS streams are fused in a fused sequence, and then, the moving object detection is applied to detect foreground regions. The fused image must represent the present information in the input images in a single signal. Therefore, to perform a pixel-level fusion successfully, all input images must be exactly spatially registered, so that all pixel positions of all the input images must correspond to the same location in the real world. In the context of a moving object detection application, some works fuse the segmentation results (foreground regions) and they consider this fusion as a mid-level fusion [52–54]. Indeed, a segmentation of input images in moving objects and background regions is performed; then, the regions of moving objects are merged either by pixel-fusion techniques [52, 53] or by a simple logical AND or OR between the masks or edges of the foreground regions [52].

Medium-level fusion In the medium-level fusion, also called feature-level fusion, the extracted features from each raw image are fused. This fusion would be obtained in the module which achieves the feature extraction and feature selection operations. Therefore, it could be achieved in two ways: between the two modules of feature extraction and feature selection or after both of them [50]. Since, one of the essential goals of fusion is to preserve the image features, the feature-level methods are able to yield subjectively better fused images than pixel fusion techniques [2].

High-level fusion In the high-level fusion, the fusion is applied either at score level or at decision level. In the score fusion, multiple classifiers produce a set of scores which represent the probabilities that one object belongs to different possible classes. These scores can be combined by a weighted parameter in order to obtain a new score which is then used to make the final decision [50]. In the decision-fusion, the classifiers are applied independently to each sensor output, and then, these decisions are combined to make a final decision. Due to the fact that decision-level fusion methods rely on the object recognition by all sensors, if there is an error in the recognition of objects from one of the sensors, the error will be transferred to the output fused result [2].

The choice of the fusion level depends on the nature of the handled application. As aforesaid, the medium-level and the high-level fusions can be treated only in the second step of our video surveillance application which is the classification of moving objects. In our context, we aim to improve the quality of moving object detection. Thus, we will be interested in low-level fusion techniques. Moreover, at this fusion level, some generic requirements are imposed on the fusion result [55]: The fusion process should preserve all relevant information of the input imagery on the composite image; it should not introduce any artifacts or inconsistencies; and the fusion process should be shift and rotational invariant. In the low-level fusion, the fusion is applied either at pixel level or at region level. In pixel fusion, we can classify the most common techniques [2, 56] into two major categories: weighted averaging-based methods and multiscale transform-based methods. In region fusion, the masks or edges of the foreground regions are merged either by pixel-fusion techniques or by a simple fusion rules.

3.2 Low-level pixel-based fusion

The low-level pixel-based fusion techniques consist in fuse the input images on a fusing image before applying any information extraction algorithm. The most widely used techniques are classified into two major categories: Weighted averaging-based methods, that are the simple averaging technique (S-Avg) and principal components analysis (PCA), and multiscale transform-based methods which are the Laplacian pyramid (*LP*), ratio of low-pass pyramid (*RoLP*), contrast pyramid (*CP*), filter-subtract-decimate pyramid (*FSD*), gradient pyramid (*GP*), discrete wavelet transform (DWT) and shift invariant discrete wavelet transform (SIDWT) techniques. To better present the different pixel-based fusion techniques, we introduce an example of IR–VIS fusion result of each technique in Fig. 1. These techniques will be detailed in the following subsection.

3.2.1 Techniques based on weighted averaging

The simplest image fusion method is to take the average of the source images (Eq. 1):

$$U_f(x,y) = \sum_{i=1}^{n} \alpha_i \, U_i(x,y) \tag{1}$$

Where $U_i(x,y)$ are the input images, α_i are the scalar weights with $\sum_{i=1}^{n} \alpha_i = 1$ and n is the number of input images. The difference between the techniques based on weighted averaging resides in the calculation method of the weights of the input images.

A. Simple Averaging technique (S-Avg) The average image of the input images is obtained with weights $\alpha_i = 1/n$. This method is the most intuitive and easiest to implement. However, all input images participate equivalently in the output image, without considering their information content. For this reason, it is considered a poor choice for the image fusion.

B. Principal Components Analysis (PCA) The optimal weighting coefficients, with respect to information content and redundant information removal, can be determined by the principal component analysis (PCA). PCA is performed on the covariance matrix C defined by Eq. 2.

$$C_{i,j} = cov(X_i, X_j) = \langle (x_i - \mu_i)(x_j - \mu_j) \rangle \tag{2}$$

Where X_i is a vector that contains all the intensities of an image, μ_i is the average of these values and $\langle \ \rangle$ is the scalar product. The covariance matrix is a squared one (n^2) where each index i, j runs through the set of the images (n). Thereafter, the eigenvalues and eigenvectors of the covariance matrix are computed. Then, the weightings for each input frame are obtained by the normalized components of the eigenvector ($\overrightarrow{V}$) corresponding to the largest eigenvalue [57] (Eq. 3):

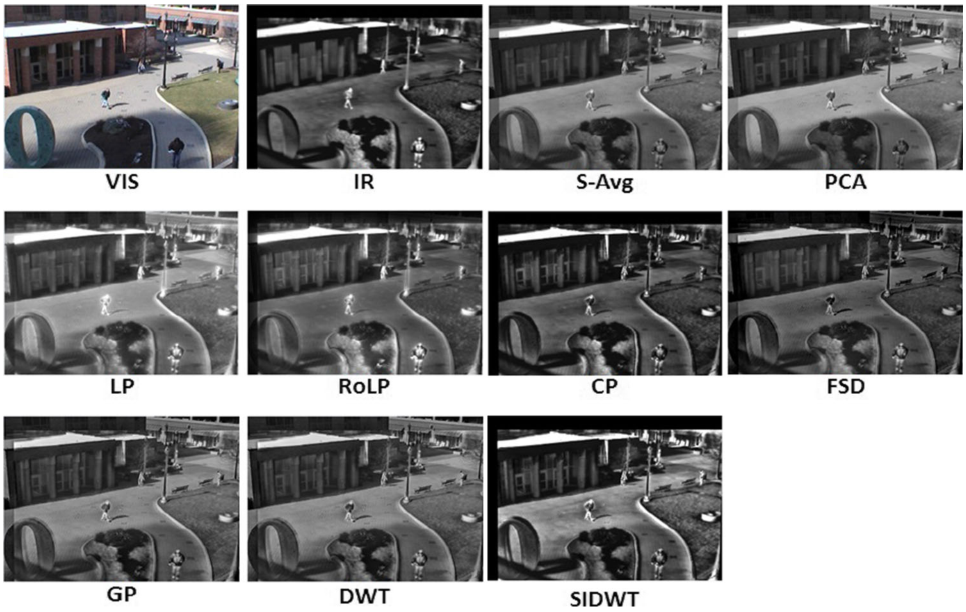


Fig. 1 The different techniques of low-level pixel-based fusion illustrated on an example of IR and VIS frames

$$\alpha_i = \frac{V_i}{\sum_{i=1}^{n} V_i} \quad (3)$$

3.2.2 Techniques based on multiscale transformation

The multiscale transformation is very useful for extracting the salient features of images. To this end, the multiscale transformation methods are the most used in image fusion. At first, they were proposed for the fusion of images with different resolutions, in the context of satellite imagery. Then, they were used for the fusion of images with the same resolution for other fusion applications. The most commonly used multiscale decomposition fusion techniques are the pyramid transforms and wavelet transforms.

A. Pyramidal Transformation techniques A pyramid structure is an efficient organization methodology for implementing a multiscale representation and computation [58]. A pyramid transform is used to represent the original image into a set of sub-images with different spatial resolutions which together represent the original image. The term "pyramid" comes from the fact that the decreasing process of the resolution engenders more often a progressive decrease in the size of sub-images. Therefore, the hierarchical structure obtained is similar to a pyramid whose stages are formed by the different levels of representation of the original image. Several techniques are used for the decomposition of an image into a pyramid representation. The most commonly used are the Laplacian pyramid (*LP*), ratio of low-pass pyramid (*RoLP*), contrast pyramid (*CP*), filter-subtract-decimate pyramid (*FSD*) and gradient pyramid (*GP*) [59–61].

B. Wavelet Transformation techniques The techniques based on wavelet transformation provide a hierarchical decomposition of a signal or an image, where each level corresponds to a higher resolution or a lower frequency band [62, 63]. The wavelet transformation is similar to the pyramid transformation, but it provides an image representation which has fewer artifacts caused by the contrast inversion, a higher signal/noise ratio and an improved perception in the merged image [54]. The term discrete wavelet transform (DWT) is a general term, involving several different methods. These methods are the most used for the task of image fusion[60]. The discrete wavelet transform refers to discrete sets of dilation and translation factors and discrete sampling of the signal.

- *Discrete wavelet transform (DWT)* is currently widely used [64, 65] for image fusion due to its facility of application and its quality of results. In practice, the DWT of an image is performed by image convolution operations with low-pass and high-pass filters for each level, followed by a sub-sampling operation.
- *Shift invariant discrete wavelet transform (SIDWT)* The fusion of image sequences based on DWT are not shift invariant, which leads to unstable and flickering fusion results. For this reason, the image sequences fusion should not be dependent on the location of an object in the image and the fusion results should be stable and consistent with the input sequences [66]. For these reasons, the SIDWT idea is to eliminate sampling, thus implicitly the interpolation in the inverse transformation to make the DWT shift invariant.

3.3 Low-level region-based fusion

In a general context, the region-fusion methods are considered as a combination of the low-level and the medium-level fusion. The basic idea is to segment an image in some way to produce a set of regions in a first place. Then, various properties of these regions can be calculated and used to determine which features from which images are included in the fused image [53, 54]. Xiao et al. [53] have proposed a region-based fusion scheme for the fusion of VIS and IR image sequences. First, both the VIS and IR sequences are enhanced using a pre-processing operator. Then, each frame of the source sequences is transformed using a multiresolution method. Simultaneously, the frames are segmented into object and background regions using a target detection method. Different fusion rules are adopted in target and background regions. Finally, the fused coefficients belonging to each region are combined, and the fused frames are reconstructed using the corresponding inverse transform. Likewise, in [54] a multiscale transformation based on wavelet transform is used to segment the features of the input images, either jointly or separately, to produce a region map. Then, the characteristics of each region are calculated and a region-based approach is used to fuse the images, region-by-region, in the wavelet domain. The regions-fusion process has the advantage of being more robust by including actual features extracted in the fused image and avoiding some of the well-known problems in pixel fusion such as blurring effects and high sensitivity to noise and mis-registration [54]. However, most of these region-fusion methods are designed for still image fusion, and each frame of each source sequence is processed individually in image sequences. These methods do not take full advantage of the wealth of inter-frame-information within the source sequences [53].

In the context of moving objects detection applications, the concept of regions-fusion methods is to fuse the foreground regions extracted by a segmentation algorithm [22, 52]. Firstly, a segmentation of input images in moving objects and background regions is performed. Then, the

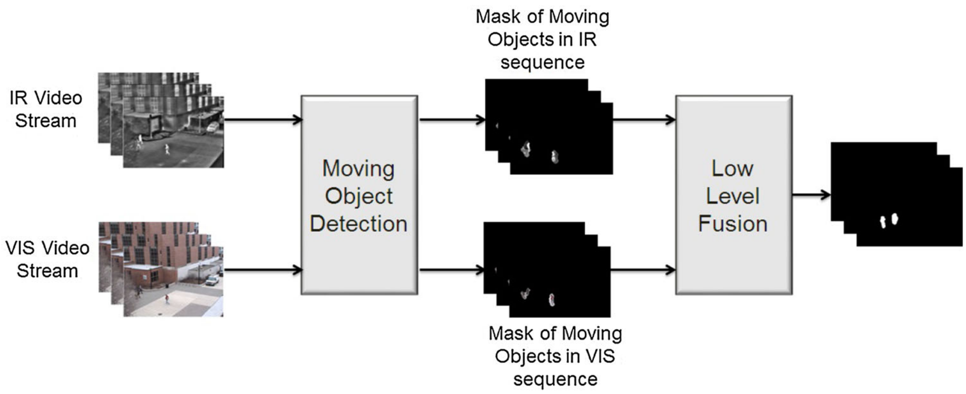

Mask of Moving Objects in IR sequence
IR Video Stream
VIS Video Stream
Moving Object Detection
Low Level Fusion
Mask of Moving Objects in VIS sequence

Fig. 2 Process of fusion of IR and VIS spectra for a moving object detection

regions of moving objects are merged either by one of the pixel-fusion techniques especially those based on multiscale transformation [52, 53] or by simple fusion rules such as a logical AND or OR between the binary masks or edges of the foreground regions [52].

3.4 Discussion

Several multisensor fusion techniques were proposed in the literature. These techniques are classified into three main levels based on the processing level where the fusion takes place. They are: low-level fusion, medium-level fusion and high-level fusion. As we aim to improve the quality of moving object detection, we are interested in the low-level fusion. At this level, the fusion can be obtained at the pixel level before applying the moving object detection or at the region level after the foreground segmentation. In the first one, there are many techniques proposed in the literature, such as weighted averaging techniques, multiscale transformation techniques which were detailed in the previous section. In the second one, the regions of moving objects are merged either by pixel-fusion techniques or by a simple fusion rules between the masks or edges of the foreground regions. In order to prove the performance of these fusion techniques and to identify the most appropriate technique for our approach, a series of experiments were carried out to evaluate the effect of the low-level fusion techniques on moving object detection results in VIS and IR spectra. These experiments are presented in the forthcoming Sect. 5.

4 Proposed method

4.1 Process

Our main contribution is to perform an accurate moving object detection all over the day (morning, afternoon and night) for particular hot objects such as people and vehicles. We propose to fuse the VIS and IR spectra to solve intricate situations and to benefit from the complementarity of the information of the two sensors. The process of our proposed method consists of two main stages: The moving object detection and the low-level fusion of IR and VIS spectra (*cf.* Fig. 2). Firstly, the study of the different categories of moving object detection methods in the IR and VIS spectra allowed us to adopt a background modeling method which is well suited to our objectives. The basic idea is to integrate the principle of inter-frame difference in the background modeling stage [67]. Thereafter, the moving object detection results in each spectrum were merged by our low-level fusion method detailed below. The effectiveness of our proposed method for a moving object detection and a low-level fusion of VIS and IR spectra was proven by the results of three sets of experiments detailed in Sect. 5.

4.2 Moving object detection

For the detection of a moving object, we adopted a method based on background modeling with a dynamic matrix and spatio-temporal analyses of scenes [67]. The representation of the background model can be achieved in two ways: recursive or non-recursive. In the recursive representation, only one model of the background is recursively updated. On the other hand, the non-recursive representation relies on a model produced from a frame buffer. In our work, we have opted for a recursive representation of the background. Such representation has low spatial and temporal complexity. In addition, it adapts quickly to sudden and gradual changes in the background. The detection of moving objects based on background modeling is generally based on four principle steps: (1) *Initialization of Background Model*, (2) *Update of Background Model*, (3) *Subtraction / Comparison* and the (4) *Post-processing* step which are shown in Fig. 3. The method adopted for each step of this process is described in subsections.

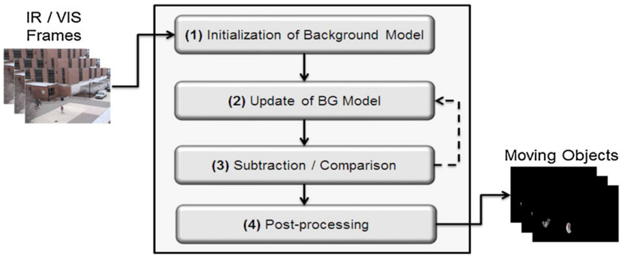


Fig. 3 Process of the adopted method for the moving object detection

(1) Initialization of background model In the initialization step, the background model can be built either online in the presence of moving objects or offline during a learning phase in the absence of moving objects. In our method, the initialization step was achieved online to benefit from such advantages as flexibility in terms of the observed scene, fast implementation and independence of the user. Indeed, three frames are used for the initialization of the background model. Knowing that in a sequence of frames obtained using a fixed camera, the moving pixels of the scene appear from one frame to another in different positions (slight differences between the values of pixels in these frames). The initial model is obtained by relying mainly on the frame at a given moment during which the non-moving pixels values will be retained in the model, while the values of the background pixels hidden by the moving pixels of this frame are approximated based on the slight differences between the pixel values of the other two frames.

(2) Update of background model The updating of the background model can be either selective where a limited number of pixels of the background are updated or non-selective where all the pixels of the model are updated. The selective update is more interesting for a better performance. However, the success of such technique depends on the candidate pixels selected for the update. Thus, our background model was updated to the changes in background pixels by a selective technique which added the pixel values to the model only if it was classified as a background pixel with significant changes. The selective update stage is based on a dynamic matrix. Indeed, the process to build the dynamic matrix operates at the frame and pixel levels. For each input frame, a pixel state card obtained by inter-frame differences based on a Dynamic Spatio-Temporal Entropy Image (DSTEI) was used to make a decision in the dynamic matrix update.

(3) Subtraction/comparison As shown in Fig. 3, the update of background model is followed by a pixel classification step moving pixel or background pixel to obtain a binary map for each pixel indicating its status. In fact, this classification was performed by subtracting each new frame from the background model built relying on an adaptive threshold, grouping the connected moving pixels in blobs and refining their shapes.

(4) Post-processing The post-processing step is performed to finalize the detection by eliminating noise and holes from the moving areas and removing uninteresting moving regions.

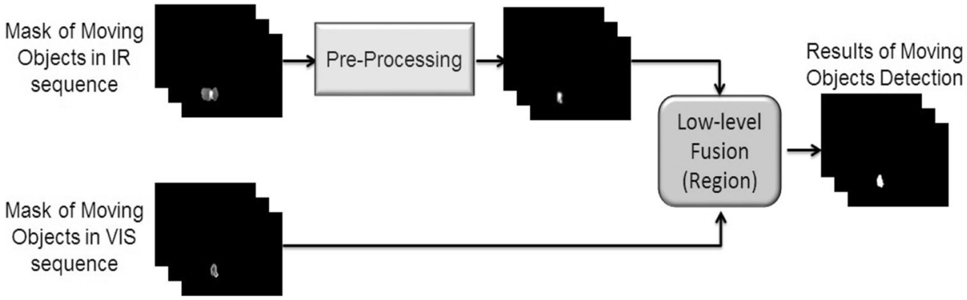
4.3 Low-level fusion

As our objective was to get the best results of moving object detection benefiting from both IR and VIS spectra, we proposed a low-level fusion method whose process is shown in Fig. 4. Firstly, we started by a pre-processing step to improve the results of a moving object detection in the IR spectrum. Secondly, we performed a fusion of the foreground regions extracted from IR and VIS spectra to assess the effect of the region-level fusion on our results of a moving object detection.

4.3.1 Pre-processing

The aim of the pre-processing step was to improve the results of foreground regions detection on the IR spectrum. In fact, one of the most relevant characteristics of the IR imagery is the hotter or colder halo surrounding the objects compared to the background. Figure 5 shows the halo phenomenon where a warm pedestrian is surrounded by a darker region than the expected background. So the halo surrounding a person would also be detected as part of the foreground region, which decreases the results of a moving object detection. Consequently, we proposed an automatic thresholding in each Bounding Box (BB) detected to

Fig. 4 Process of the proposed method for low-level fusion


Mask of Moving Objects in IR sequence
Pre-Processing
Results of Moving Objects Detection
Low-level Fusion (Region)
Mask of Moving Objects in VIS sequence

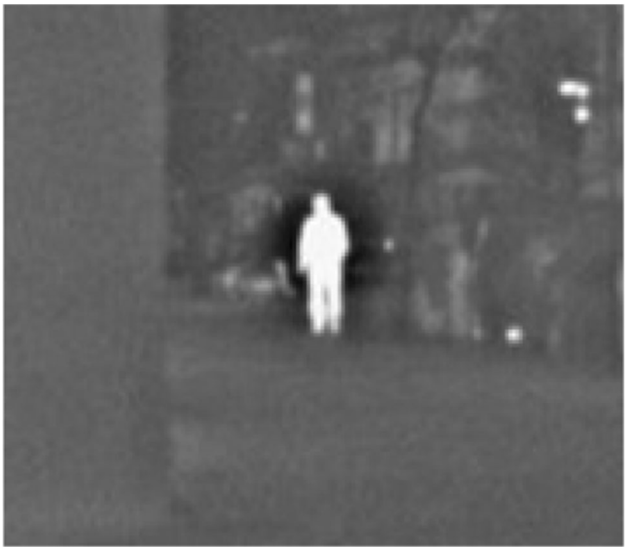


Fig. 5 Example of halo surrounding pedestrian in a typical winter day

extract the bright pixels that belong to the pedestrian and removed the halo. The originality of this step is the adaptive threshold calculated for each BB. First, we calculated the sum of each BB column. In fact, the column that has the maximum of pixels belongs to pedestrian and would have the maximum amount. Then, the BB adaptive threshold value was calculated that is the average value of the selected column. Finally, the adaptive threshold was applied to the BB in order to select only pixels corresponding to warm objects. Therefore, we used morphological operations to fill holes inside the blobs. The pseudocode of our pre-processing algorithm is presented in Algorithm 1.

4.3.2 Low-level region-based fusion

The goal of our proposed method was to achieve a better detection of moving objects using two different types of cameras one in the IR spectrum and another in VIS spectrum. As mentioned previously in the state of the art, at this processing level, the fusion of the IR and VIS spectra was performed with low-level fusion techniques either pixel-based or region-based. The choice of the most appropriate low-level fusion technique for our proposed method is based on a thorough experimental study detailed in Sect. 5. The experimental results showed that a low-level region-based fusion improved the moving object detection results better than the pixel-fusion techniques. In fact, the low-level region-based fusion has the advantage of preserving the image quality of the two spectra compared to the pixel-based techniques that may alter the fused image. Our low-level region-based fusion method was achieved by a logical AND between the binary masks of foreground regions detected separately in the IR and the VIS spectra. Finally, the resulting masks of moving objects were used in a higher processing level such as the feature extraction for the classification of moving objects.

5 Experimentations

In order to study the performance of the proposed method of a moving object detection and the different techniques of low-level fusion, three series of experiments were carried out. The first one was used to check whether the various pixel-fusion techniques would improve the performance of the moving object detection, rather than using either of the

Algorithm 1 Pre-processing step

```
Input: Bounding Box BB (m,n) of moving object in IR spectrum
Output: Binary mask BW of the Bounding Box
  (a) Calculate the sum of each column of the BB in the vector V
  (b) Look for the column that has the maximum sum from the vector V: Cmax
  $max \leftarrow V(1)$
  $p \leftarrow 1$
  for $i = 2$ to $n$ do
    if $V(i) >= max$ then
      $max \leftarrow V(i)$
      $p \leftarrow i$
    end if
  end for
  $Cmax \leftarrow BB(:, p)$
  (c) Calculate the adaptive threshold Th
  $Th = mean(Cmax)$
  (d) Automatic thresholding of each pixel (i,j) of the BB
  if $BB(i, j) >= Th$ then
    $BW(i, j) \leftarrow 1$
  else
    $BW(i, j) \leftarrow 0$
  end if
```

Table 1 Details of the dataset

Dataset	IR/VIS Sequences	Number of Frames	Moving objects class	Weather Conditions	Challenges
Scene 1	Seq-1	2107	Person	Sunny day	Shadow cast by moving objects and a group of clouds in VIS spectrum
	Seq-2	1201	Person		
	Seq-3	3399	Person		
Scene 2	Seq-4	3011	Person	Cloudy day	Halos surrounding moving objects in IR spectrum
	Seq-5	4061	Person		
	Seq-6	3303	Person		
INO	GF	1482	Car and person	Daytime scene	Moving objects occlusion
	PS	2941	Car and person	Cloudy day	
	PE	820	Car and person	Evening scene	Low contrast of moving objects
	CP	240	Person	Sunny day	

two sensors independently. In the second series, we evaluated firstly the relevance of the pre-processing step applied on the results of the detection of moving objects in the IR spectrum, and secondly, the efficiency of our low-level fusion method against the two independent sensors was assessed. Finally, a comparison of our method with other potential methods was achieved. These methods are based on background subtraction for the detection of moving objects in IR and VIS spectra and include low-level region-based fusion resulting in a rich test field of fusion techniques. Before presenting the results of these series of experiments, we described the database and the used validation techniques.

5.1 Dataset description

To evaluate quantitatively and qualitatively the proposed method of a moving object detection and the different low-level fusion techniques of IR and VIS spectra, we have selected two popular databases shown in Table 1 . The first dataset is the OTCBVS Benchmark Dataset[1], namely “OSU Color-Thermal Database” consisting of six challenging thermal/color video sequence pairs recorded from two different locations at different times-of-day, with different camera gain and level settings. The thermal sequences were captured using a Raytheon 300D ferroelectric BST thermal sensor core, and a Sony TRV87 Handycam was used to capture the color sequences. The image sizes were half-resolution at 320 × 240 pixels. The sequences were recorded on the Ohio State University campus during the months of February and March 2005, and show several people, some in groups, moving through the scene. Sequences 1, 2 and 3 (scene 1) contain regions of dark shadows cast by the buildings in the background. There are also frequent and abrupt illumination changes in various parts of the observed scene caused by a group of clouds passing in the sky. The images of Sequences 4, 5 and 6 (scene 2) were captured on a cloudy day, with fairly constant illumination and soft/diffuse shadows. To incorporate variations in the thermal domain, the gain/level settings on the thermal camera were varied across the sequences [52]. A manual segmentation of the moving objects regions was performed in 80 images both in VIS and IR sequences from this dataset for the quantitative evaluation. Indeed, this segmentation has been neatly carried out by two persons of our laboratory. The results of the hand segmentation of each pair of images by each person were fused using a logical AND to guarantee a precise segmentation of foreground regions. The INO Video Analytics dataset[2] made available with their ground-truth (GT) frames. In fact, the infrared sequences are recorded with ThermoVision A10 camera which has a VO2-based microbolometer detector array of 164 × 128 pixels and a sensitivity of about 80 mK, whereas a Marlin F33C CCD camera (640 × 480 pixels with Bayer filter) was used as color sensor. In the test dataset, we use four thermal/color video sequences which are the Group-Fight (GF), the Parking-Snow (PS), the Parking-Evening (PE) and the Close-Person (CP) sequence. These sequences are recorded in various locations and covering different weather conditions. This dataset, illustrated in Fig. 6 with examples of IR, VIS and ground-truth images from each sequence, constitutes a rich test field to validate our proposed method.

5.2 Validation techniques

For the performance evaluation, we have used some evaluation metrics to judge the effectiveness of the different techniques of low-level fusion on the detection of moving

[1] http://vcipl-okstate.org/pbvs/bench/.

[2] http://www.ino.ca/en/video-analytics-dataset/.

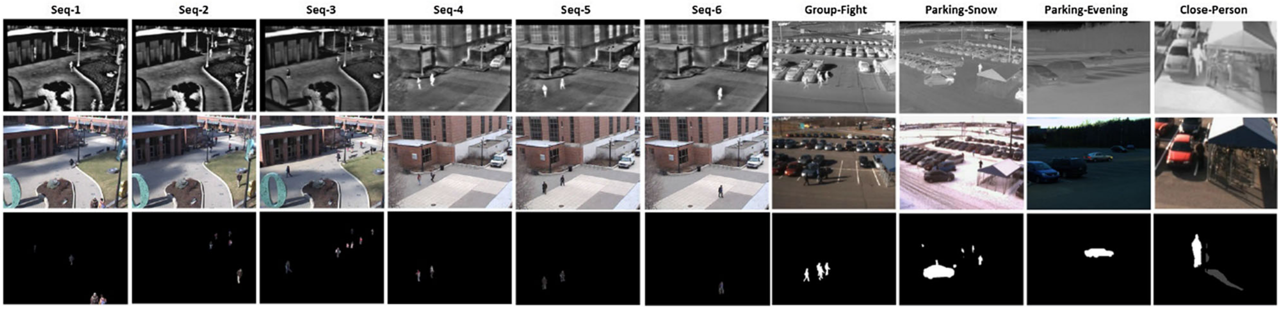
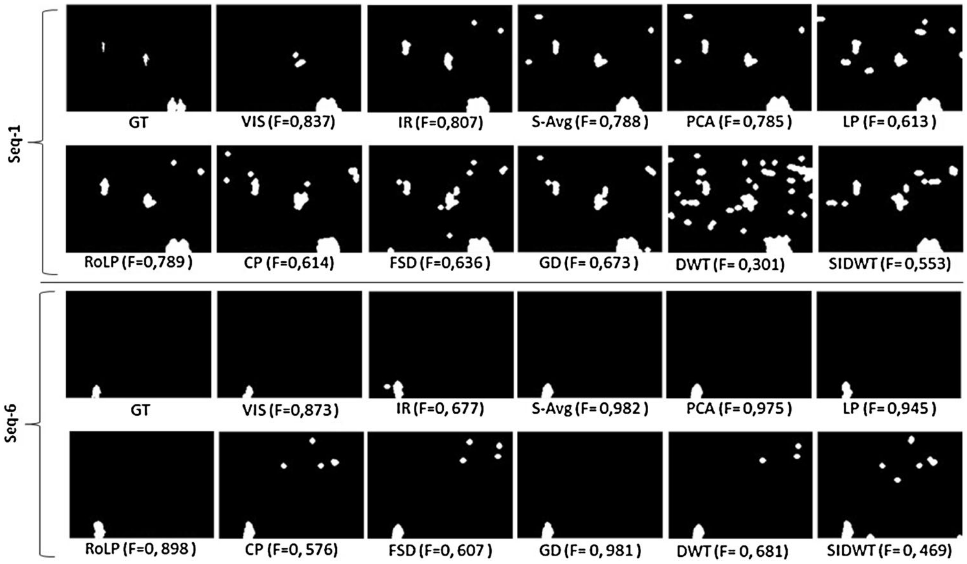


Fig. 6 Examples of IR, VIS and ground-truth images

Fig. 7 Visual comparison of the detection results on two images, one from the *Seq-1* (frame number 217) and one from the *Seq-6* (frame number 443), in the IR, VIS spectra and the different fused images



objects. In most of the studies, researchers work on the construction of a confusion matrix to evaluate the quality of the detection of the moving objects. From this confusion matrix we measured the *Recall* and *Precision* using the set of 80 manually labeled images as ground truth. The *Recall* rate is also calculated to know accurately the rate of moving object pixels that are correctly detected, while *Precision* rate refers to the rate of correct classification. We also used the F-measure [68], which is the harmonic mean of Recall and Precision. In fact, a higher F-measure value corresponds to a higher value of Recall and Precision (Eq. 4).

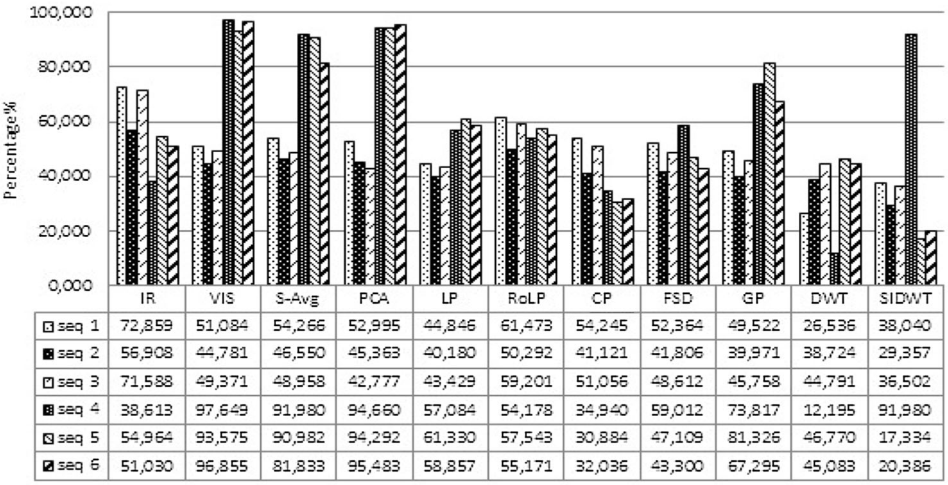
$$F = \frac{2 \times P \times R}{P + R} \tag{4}$$

5.3 Experimental results

In this section, we present the results of the three series of experiments detailed above.

5.3.1 *Experiment 1: independent sensors versus low-level pixel-based fusion*

The aim of this experiment was to evaluate the impact of low-level pixel-based fusion techniques on the results of moving object detection. The pixel-fusion techniques were applied on the input sequences to extract the fused sequences of the IR and VIS spectra. Then, the proposed method of a moving object detection was applied on the different fused sequences to evaluate the impact of pixel-level fusion techniques on the results of a moving object detection. Using all manually labeled images as a ground truth, we computed Recall, Precision and F-measure of the results of a moving object detection on the IR and VIS sequences and the fused sequences by S-Avg, PCA, LP, RoLP, CP, FSD, GP, DWT and SIDWT techniques. Figure 7 shows an example of detection results (Seq-1 and Seq-6) in IR, VIS images and fused images by the low-level fusion techniques. For each frame, we calculated the F-measure value (F). As is clear in these visual results, the



	IR	VIS	S-Avg	PCA	LP	RoLP	CP	FSD	GP	DWT	SIDWT
seq 1	72,859	51,084	54,266	52,995	44,846	61,473	54,245	52,364	49,522	26,536	38,040
seq 2	56,908	44,781	46,550	45,363	40,180	50,292	41,121	41,806	39,971	38,724	29,357
seq 3	71,588	49,371	48,958	42,777	43,429	59,201	51,056	48,612	45,758	44,791	36,502
seq 4	38,613	97,649	91,980	94,660	57,084	54,178	34,940	59,012	73,817	12,195	91,980
seq 5	54,964	93,575	90,982	94,292	61,330	57,543	30,884	47,109	81,326	46,770	17,334
seq 6	51,030	96,855	81,833	95,483	58,857	55,171	32,036	43,300	67,295	45,083	20,386

Fig. 8 Quantitative results (F-measure) of object detection on IR, VIS sequences and the fused sequences by S-Avg, PCA, LP, RoLP, CP, FSD, GP, DWT and SIDWT

moving object detection results varied between the two examples. In fact, for Seq-1 the detection results in the fused images have degraded compared to those in the two spectra independently. Whereas for the second example, the detection result in the fused images by s-Avg, PCA, LP, RoLP and GD techniques were improved as compared to the IR and VIS spectra. We performed a quantitative evaluation of the detection method on IR, VIS sequences and of the fused sequences by the low-level pixel-based fusion techniques. From the results, presented in Fig. 8 in terms of F-measures average of each sequence, we can conclude that the pixel-fusion techniques have not improved the detection results. Their results depended on the challenges present in the scenes and varied from one sequence to another. As it is clear in the F-measure values, the detection results on either of the sensors independently were almost always better than those in the fused sequences. Indeed, in the Sequences 1, 2 and 3 of scene 1, the IR spectrum has recorded the best rates of moving objects detection (respectively 72,859%, 56,908% and 71,588%) compared to the other results. However, in the Sequences 4, 5 and 6 of scene 2, the moving object detection results had the best rates in the VIS spectrum. Unless, in sequence 5, the detection results in the fused images by PCA low-level fusion technique were slightly better than the VIS spectrum (94,292% against 93.575%). Therefore, we can conclude that a fusion of the two IR and VIS images can degrade the results of an object detection instead of improving them.

5.3.2 *Experiment 2: independent sensors versus low-level region-based fusion*

The aim of this experiment was to evaluate firstly the relevance of the pre-processing step applied on the results of the detection of moving objects in the IR spectrum and secondly, the efficiency of our low-level fusion method against the two independent sensors. Figure 9 shows examples of the detection results extracted from each IR sequences of our dataset, (a) before pre-processing and (b) after pre-processing. In the sequences of scene 1, the detection results are slightly degraded. Indeed, these sequences were recorded during a sunny day so there were many hot areas and the objects were not well contrasted from the background. On the other hand, in the sequences of scene 2, the detection results have greatly improved. In fact, these sequences were captured on a cloudy day and the pedestrians were very bright from the background. So, in each BB halo surrounding the pedestrians were successfully eliminated. In the sequences of the INO dataset, the pre-processing step has improved the detection results in the IR spectrum except in the PE sequence. In this latter, the moving objects have low contrast with the background mostly for the vehicle class. In the PE example of Fig. 9, our pre-processing has created a hole inside the detected vehicle because it was not well illuminated. In the PS and CP sequences the detection results were improved as shown in the examples presented in Fig. 9. These qualitative results are justified by the values of R, P and F of the detection results calculated using the ground-truth images of each sequence. In fact, as shown in Table 2, the F-measures average has been improved from 66.019% to 72.978% by adding a pre-processing step to the detection of moving objects in the IR spectrum. Although our pre-processing step has reduced somewhat the recall rates in some sequences (from 80,072% to 75,013% in GF and from 81.75% to 75.82% in PE), the precision rates were obviously improved (from 81,906% to 91,808% in GF and from 74,419% to 79,51% in PS). This is due to the fact that some parts of the moving objects, mainly the car class, lack brightness. In this case, our pre-processing step will cause some holes in these foreground regions, which degrades the

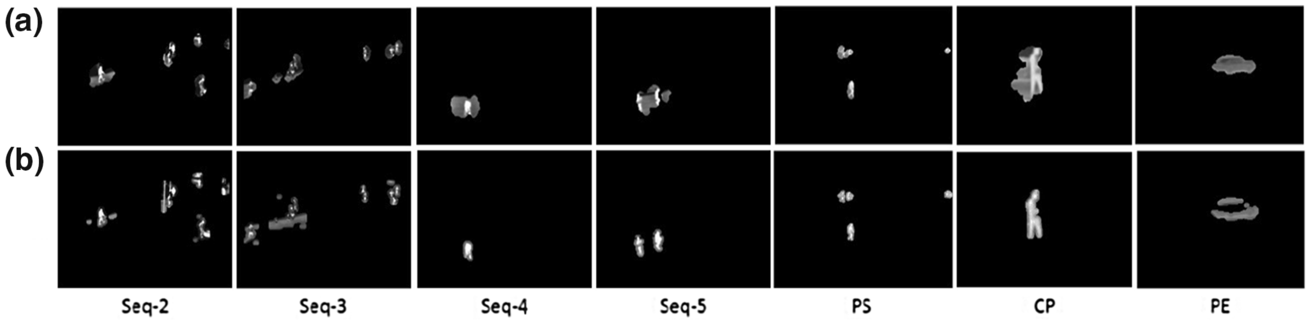


Fig. 9 Qualitative results of a moving object detection before pre-processing (**a**) and after pre-processing (**b**) in each IR sequence

Table 2 Comparison of the performance of our method in IR spectrum without and with the pre-processing step

		Seq-1	Seq-2	Seq-3	Seq-4	Seq-5	Seq-6	GF	PS	PE	CP	AVG
MOD in IR spectrum	R	96.173	95.590	92.679	79.039	86.389	85.372	80.072	72.418	81.75	70.65	84.013
	P	60.934	42.100	59.452	26.546	42.664	37.320	81.906	74.419	92.223	75.876	59.344
	F	72.859	56.908	71.588	38.613	54.964	51.030	80.979	73.405	86.671	73.17	66.019
*MOD + Pre-processing	R	95.920	98.045	94.951	100.00	99.826	97.856	75.013	79.058	75.82	75.776	89.227
	P	42.575	29.216	48.486	77.469	76.699	78.918	91.808	79.51	82.248	74.957	68.189
	F	57.175	43.694	61.976	84.292	82.836	83.690	82.565	79.283	78.903	75.3643	72.978

recall rates. The lowering in all rates of PE sequence is due to the low contrast and brightness of moving objects in IR images. It is worth noting that our pre-processing step does not influence the processing time of our method because it has an extra cost of 0.45% in terms of computing time. The recorded time and the obtained results reveal the efficiency of the processing step without affecting the time processing. Therefore, our method provides a good compromise between processing time and efficiency.

To benefit from the complementary nature of the two sensors, a fusion of foreground regions extracted from IR and VIS spectra was carried out. Indeed, this low-level region-based fusion has the advantage of preserving the image quality of the two spectra compared to the pixel-based techniques that may alter the fused image. For this task, we combined the detection results from the two sensors by performing a simple intersection of their foreground regions using the logical AND between their binary masks. In fact, this fusion reduced the halo effect in the detection results with the IR spectrum due to the absence of this phenomenon in the VIS spectrum, on the one hand, and the moving shadow recorded in the VIS sequences on the other hand. Our corpus consists of a rich sequence set recorded in different locations under various weather conditions and covering different challenges, allowing us to validate our proposed method. Thus, the effectiveness of our low-level fusion method, including the pre-processing step and region fusion of IR and VIS spectra, was proven by the results presented in Table 3. Indeed, the average of the F-measure values has increased from 66.019% in the IR spectrum and from 74.716% in the VIS spectrum to 85.063% with the proposed low-level fusion. An example of the qualitative results of our low-level fusion results is shown in Fig. 10. In scene 1, the low-level fusion has improved more the detection results especially compared to the VIS spectrum because there were many regions of moving shadows. In sequences of scene 2, the low-level fusion has improved the detection results than the IR spectrum because there were many halos surrounding the moving objects which are eliminated by our pre-processing method. Regarding the GF, PS, PE and CP sequences of the INO dataset, our low-level fusion has improved the moving object detection results rather than use only one of the two spectra. The importance of our framework is related to the complexity of the observed scenes which present a variety of challenges. Our method has proven its potential to overcome several challenges. Figure 11 shows some examples of detection results by our proposed method under different challenges. In fact, the frames (a) number 65 and (b) number 102 of seq-1 present the illumination changes challenges where the moving object is covered by a dark shadow caused by a group of cloud passing in the sky. The frames (c–g), extracted from PE and GF sequences, show the performance of our method to detect moving objects with different sizes and different speeds (vehicle versus pedestrian). Finally, in the frames (h–j)

Table 3 Comparison of moving object detection results in the IR and VIS spectra and the fusion of the two spectra using the proposed method of low-level fusion

		Seq-1	Seq-2	Seq-3	Seq-4	Seq-5	Seq-6	GF	PS	PE	CP	AVG
IR	R	96.173	95.590	92.679	79.039	86.389	85.372	80.072	72.418	81.75	70.65	84.013
	P	60.934	42.100	59.452	26.546	42.664	37.320	81.906	74.419	92.223	75.876	59.344
	F	72.859	56.908	71.588	38.613	54.964	51.030	80.979	73.405	86.671	73.17	66.019
VIS	R	90.064	89.855	86.863	96.747	89.591	94.908	68.215	83.544	75.7	78.173	85.366
	P	39.143	35.355	39.585	98.986	99.064	99.236	98.935	96.648	47.731	92.948	74.763
	F	51.084	44.781	49.371	97.649	93.575	96.855	80.752	89.62	58.547	84.923	74.716
Low-level fusion	R	83.953	87.835	84.101	97.449	89.838	95.530	75.066	70.683	65.651	74.048	82.4154
	P	78.965	69.975	83.673	97.176	98.645	93.941	96.854	99.917	94.124	93.409	90.668
	F	80.789	76.101	82.581	97.037	93.194	93.598	84.579	82.795	77.351	82.609	85.063

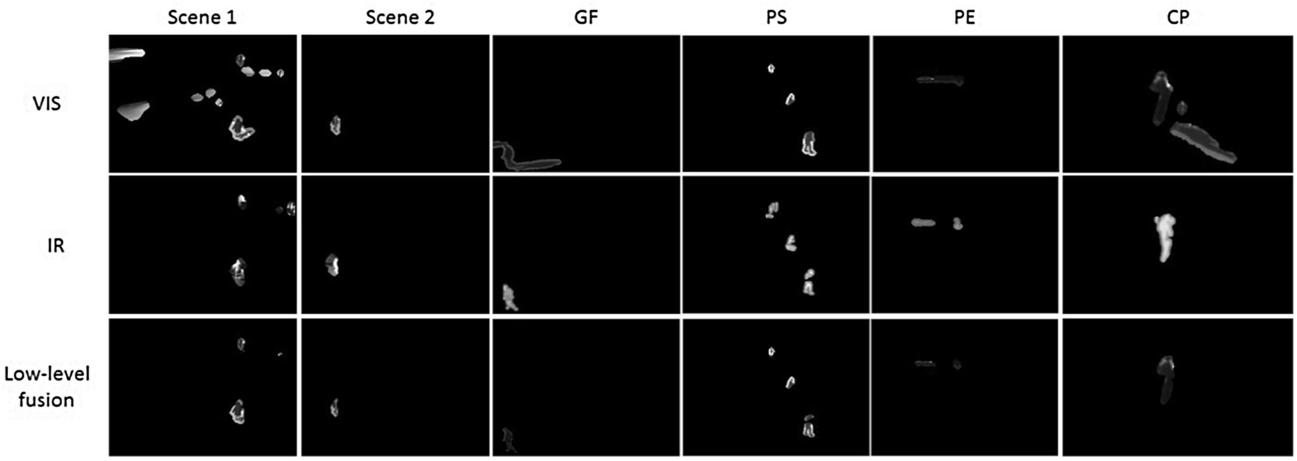


Fig. 10 Visual comparison of the detection results in the IR, VIS spectra and our low-level fusion results

there are many partial occlusions when the pedestrians were hidden by a stationary car. The set of experiments presented above has demonstrated the efficiency of our preprocessing step applied to the detection results of the IR sequences. The second novelty of this method lies in the fusion of two sensors together rather than the use of one sensor independently.

5.3.3 *Experiment 3: our proposed method versus other methods*

For our application domain, the lack of an open access to the codes, IR datasets with their ground-truth frames and detailed descriptions of algorithms hinders the elaboration of a fair comparison with several methods and on large datasets. Nevertheless, we have evaluated the performances of our method using a comparative platform that includes two well-known methods [52, 69] and three recent methods [70–72] of moving object detection in IR and VIS spectra. These methods are compared with our results separately given that each one of them has presented its results differently.

First, we start by qualitative evaluation of our detection results on the IR spectrum compared to those of T. Parag et al. [70]. In fact, they propose a moving object detection method based on a probabilistic method incorporating background subtraction method. In fact, they associate a label for each pixel which indicates whether it is a target or background pixel. The optimal label set for all the pixels of an image maximizes aposteriori distribution of label configuration given the pixel intensities. The posterior probability is factored into a conditional likelihood of the intensity values and a prior probability of label configuration. These two probabilities are computed assuming a Markov random field (MRF) on both pixel intensities and their labels. Furthermore, they enforce neighborhood dependency on both intensity values, by a simultaneous auto-regressive (SAR) model, and on labels, by an auto-logistic model. The parameters of these MRF models are learned from labeled examples. During testing, an MRF inference technique, namely Iterated Conditional Mode

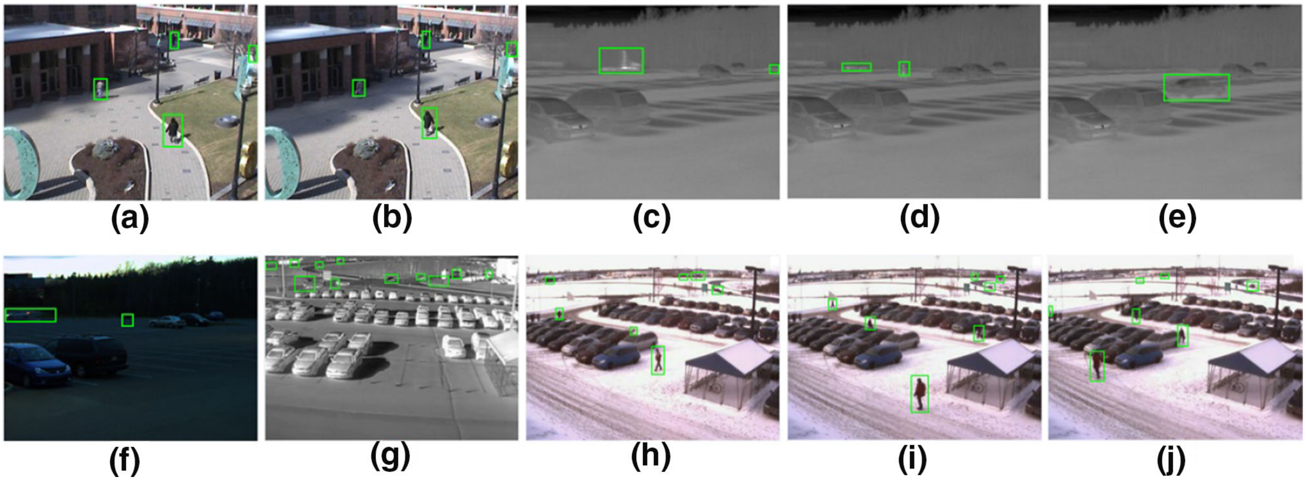
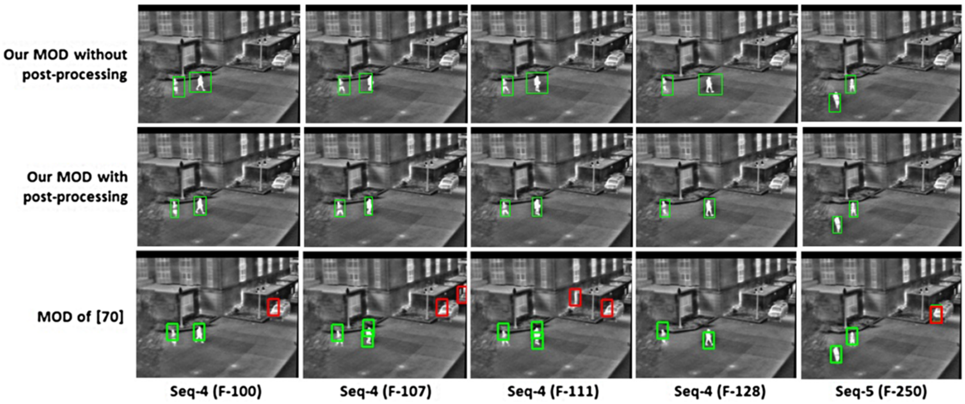


Fig. 11 Qualitative results of our moving object detection method using low-level fusion of IR–VIS spectrum facing to: **a**, **b** illumination changes, **c**–**g** object size and speed changes and **h**–**j** occlusions

Fig. 12 Comparison of moving object detection method without and with pre-processing in IR frames with the work of [70]



(ICM), produces the optimal label for each pixel. Finally, the detection performance is further improved by incorporating temporal information through the background subtraction. We present in the first and second rows of Fig. 12, respectively, our moving object detection without and with our pre-processing step on IR images of Seq-4 (frame number 100, 107, 111 and 128) and Seq-5 (frame number 250). In the last row we present the qualitative results of T. Parag et al. for the same frames which are presented in [70]. The red bounding box refers to erroneous detections. These results show the effectiveness of our method compared to those of [70] which present some errors in the detection of the moving pedestrians.

In the second set of this experimentation, we present quantitative evaluations with four other works [52, 69, 71, 72]. These methods are similar to our own as they use detection methods based on background modeling and include low-level region-based fusion of IR and VIS spectra. The comparison was carried out on the same frames and using the same metrics presented by each work. Firstly, we compare our results with two well-knowing methods [52, 69]. The first method (1), which recorded the six sequences of OTCBVS dataset used in our experiments, was introduced by Davis et al. [52]. The authors used a background subtraction in the IR spectrum to identify the initial regions of interest. Then, they identified the corresponding regions of interest in the VIS spectrum based on color and intensity information. Within each region, the input and background gradient information were combined to form a Contour Saliency Map. The binary contour fragments, obtained from the corresponding Contour Saliency Maps of the two sensors, were then fused into a single image by performing a simple union. Lastly, the contour image was flood-filled to produce silhouettes of the detected moving objects. The second method (2) was proposed by Kim et al. [69] and can handle scenes containing moving backgrounds or illumination variations. This method uses a codebook technique to build the background models independently for both spectra. It also allows capturing structural background variations due to periodic-like motion over a long period of time under limited memory. The codebook representation is efficient

Table 4 Comparison of precision (P), recall (R) and F-measures (F) values of different methods across different sequences

Methods		Seq-1	Seq-2	Seq-3	Seq-4	Seq-5	Seq-6	AVG
(1) Davis *et al.*	R	0.756	0.754	0.683	0.759	0.823	0.814	0.755
	P	0.908	0.890	0.900	0.958	0.965	0.913	0.916
	F	0.825	0.816	0.776	0.847	0.888	0.827	0.828
(2) Kim *et al.*	R	0.772	0.561	0.543	0.925	0.932	0.914	0.733
	P	0.747	0.568	0.915	0.910	0.909	0.915	0.779
	F	0.759	0.564	0.681	0.917	0.920	0.914	0.755
(Our)	R	0.840	0.878	0.841	0.974	0.898	0.955	0.898
	P	0.790	0.700	0.837	0.972	0.986	0.939	0.871
	F	0.808	0.761	0.826	0.970	0.932	0.936	0.872

in memory and speed compared to other background modeling techniques. The background subtraction results in each spectrum are fused using region-level fusion by a pixel-wise logical OR. This method was evaluated on the same dataset of the six sequences recorded by [52]. The comparison of our method with the two other methods (1) and (2), as presented in Table 4, proves the effectiveness of our proposed method based on low-level fusion of the IR and VIS sensors for moving objects detection. In fact, our method records the highest F-measure 0.872 for large number of ground-truth frames (20 higher than [52, 69]). Although only the F-measure values recorded in seq-1 and seq-2 by method (1) are better than our results, we are far better in term of recall values. In fact, the decrease in precision values is due to the strong presence of moving shadows and the noise caused by frequent illumination changes caused by a group of clouds passing in the sky. Otherwise, we are better in all other sequences and in the average F-measure of all sequences of the dataset. Our considerably higher accuracy than the competing methods show the performance of our method to deal with the challenges of both domains, namely halos, shadows and illumination changes.

In the second work, the authors of [71] have proposed a moving object detection method based on background modeling using the Gaussian mixture models and present a low-level region-based fusion method named “Late-Fusion” to merge the IR and VIS spectra. We compared our detection results with those of this work in the IR spectrum, the VIS spectrum and with their low-level fusion method. In fact, in the proposed method, they have applied the background subtraction at each modality separately, and then, they have used a book-keeping algorithm for a number of frames to check the consistency of the pixels nature, if the two foreground maps do not agree on the nature of a pixel. In other words, the pixels state remains unchanged if there is inconsistency between the two sensors. Table 5 shows a comparison between our detection results and those of [71]. In fact, we record the best rates of moving object detection in each one of the IR and VIS spectra except the visible PE and the infrared PS sequences. In the fused sequences, our low-level fusion gave better detection results in all sequences except in PE sequence due to the fact that our moving object detection method has yielded somewhat low results in the visible spectrum. In fact, this sequence was recorded in the evening and the moving objects have low contrast with background in the VIS images.

Table 5 Comparison of detection results (AVG(F-measure)) of our method and the work of [71] in the IR spectrum, VIS spectrum and with different low-level fusion methods

Methods	GF	PS	PE
VIS of [71]	0.679	0.846	0.752
Our VIS	0.808	0.896	0.585
IR of [71]	0.758	0.871	0.733
Our IR	0.826	0.793	0.789
Late fusion of [71]	0.835	0.613	0.883
Our low-level fusion	0.846	0.828	0.774

Finally, in the recent work of [72] the authors have proposed a background subtraction method which is performed separately in IR and VIS spectra and a low-level region-based fusion. In fact, after foreground segmentation, they use connected component theory and fusion rules to fuse the extracted regions in each modality to get additional information about the moving objects. Table 6 shows a comparison between the detection results of [72] and those of our proposed method, in terms of precision metric on GT images and the average precision of each sequence. In each sequence, we find in the first line, all the numbers

Table 6 Comparison of precision rates of our low-level fusion method and the work of [72]

	CP							
	146	161	181	196	201	226	236	AVG
[72]	0.982	0.989	0.968	0.964	0.951	0.941	0.921	0.959
Our fusion	0.853	0.826	0.89	1	1	1	0.994	0.938
	PS							
	751	851	1651	2251	2351	2551	2741	AVG
[72]	0.808	0.928	0.821	0.836	0.787	0.78	0.793	0.822
Our fusion	1	0.987	1	1	1	1	1	0.998
	GF							
	378	528	678	1028	1178	1278	1328	AVG
[72]	0.898	0.877	0.968	0.849	0.66	0.604	0.696	0.793
Our fusion	0.987	1	0.982	0.976	0.99	1	0.996	0.990

of the GT images select by the authors of [72] and in the following line their precision rates and those of our low-level fusion method. In the last column of each sequence, we find the precision averages of our results and those of [72]. As shown in Table 6, we have recorded the best rates in all frames of these sequences (CP, PS and GF) except in the frame number 146, 161 and 181 of Close-Person sequence. The high precision rates of our results prove the robustness of our moving object detection method by a low-level fusion of the IR–VIS spectra.

6 Conclusion

In this paper, we presented a classification of the different moving object detection methods in IR and VIS spectra and presented a state of art on fusion methods. The study of the different categories of moving object detection methods in the IR and VIS spectra has allowed us to adopt a method based on background modeling incorporating the principle of inter-frame difference in the background modeling stage. The fusion methods are classified into three main levels: low-level fusion, medium-level fusion and high-level fusion. The low-level fusion has two sub-levels: pixel-level and region-level fusions. We first demonstrated that a pixel-level fusion of the IR and VIS spectra can degrade the results of an object detection compared to the detection results on either of the sensors independently. To this end, we proposed a low-level fusion based firstly on a pre-processing step to improve the results of a moving object detection on the IR spectrum; Secondly, on a low-level region-based fusion of a foreground area. Our approach was evaluated qualitatively and quantitatively compared to other works using a rich test field consisting of ten challenging thermal/color video sequences recorded from different locations and at different times-of-day covering different weather conditions. We used a set of GT images both in VIS and IR sequences to perform a thorough quantitative evaluation based on the Precision and Recall values of the detected foreground regions. Three series of experiments were performed to prove the effectiveness of our approach. In the first experiment, we evaluated the impact of low-level pixel-based fusion techniques on the results of moving object detection. The second experiment proved the efficiency of our low-level fusion method on the detection results when using thermal and visible imagery together, over using either domain independently. Finally, we compared our low-level fusion results with two well-known methods and three recent methods from the literature which propose moving object detection in IR and VIS spectra using low-level fusion methods. Our best rates recorded throughout the various experiments presented demonstrate the robustness and efficiency of our method for a moving object detection using a low-level region-based fusion of IR and VIS spectra. Our future perspective will examine the semantic classification of the detected moving objects in the IR and VIS spectra using our promising results while studying the medium-level and the high-level fusion.

References

1. Pavlidis I, Morellas V, Tsiamyrtzis P, Harp S (2001) Urban surveillance systems: from the laboratory to the commercial world. Proc IEEE 89(10):1478–1497
2. Zin TT, Takahashi H, Toriu T, Hama H (2011) Fusion of infrared and visible images for robust person detection. In: Ukimura O (ed) Image fusion. InTech, Rijeka

3. Conaire CO, Cooke E, O'Connor N, Murphy N, Smearson A (2005) Background modelling in infrared and visible spectrum video for people tracking. In: Proceedings of international conference on computer vision and pattern recognition, San Diego, California, pp 20–20
4. Prajapati D, Galiyawala HJ (2015) A review on moving object detection and tracking. Int J Comput Appl 5(3):168–175
5. Joshi KA, Thakore DG (2012) A survey on moving object detection and tracking in video surveillance system. Int J Soft Comput Eng 2(3):44–48
6. Xu Y, Dong J, Zhang B, Xu D (2016) Background modeling methods in video analysis: a review and comparative evaluation. CAAI Trans Intell Technol 1(1):43–60
7. Cheng YH, Wang J (2014) A motion image detection method based on the inter-frame difference method. Appl Mech Mater 490–491:1283–1286
8. Arvanitidou MG, Tok M, Glantz A, Krutz A, Sikora T (2013) Motion-based object segmentation using hysteresis and bidirectional inter-frame change detection in sequences with moving camera. Image Commun J 28(10):1420–1434
9. Zhen Y, Yanping C (2009) A real-time motion detection algorithm for traffic monitoring systems based on consecutive temporal difference. In: Proceedings of 7th Asian control conference, Hong Kong, pp 1594–1599
10. Xin W, Gaolue L (2011) Fusion algorithm for infrared–visual image sequences. In: Proceedings of the 6th International Conference on Image and Graphics, Hefei, Anhui, pp 244–248
11. Fernandez-Caballero A, Castillo JC, Martinez-Cantos J, Martinez-Tomas R (2010) Optical flow or image subtraction in human detection from infrared camera on mobile robot. Robot Auton Syst 58(12):1273–1281
12. Jain R, Nagel H-H (1979) On the analysis of accumulative difference pictures from image sequences of real world scenes. IEEE Trans Pattern Anal Mach Intell 1(2):206–214
13. Lillestrand RL (2006) Techniques for change detection. IEEE Trans Comput C–21(7):654–659
14. Asli RN, Zavaraki MM (2016) Fast-optimized object detection in dynamic scenes using efficient background weighting. Int J Hybrid Inf Technol 9(3):11–22
15. Pang Y, Ye L, Li X, Pan J (2015) Moving object detection in video using saliency map and subspace learning. In: IEEE Transactions on Circuits Systems for Video Technology, pp 4321–4330
16. Hou AL, Guo JL, Wang CJ, Wu L, Li F (2013) Abnormal behavior recognition based on trajectory feature and regional optical flow. In: Proceedings of the 7th international conference on image and graphics, Qingdao, pp 643–649
17. Jian-Ping T, Xiao-lan L, Jun L (2016) Moving object detection and identification method based on vision. Int J Secur Appl 10(3):101–110
18. Ke H (2016) Moving object detection research based on background image set and sparse analysis. J Softw Eng 10(1):66–77
19. Akula A, Khanna N, Ghosh R, Kumar S, Das A, Sardana HK (2014) Adaptive contour-based statistical background subtraction method for moving target detection in infrared video sequences. Infrared Phys Technol 63:103–109
20. Chen BW, Liu SL (2014) Infrared target detection based on temporal–spatial domain fusion. Adv Mater Res 1044–1045:1186–1189
21. Bondzulic B, Belgrade MA, Petrovic V (2008) Multisensor background extraction and updating for moving target detection. In: Proceedings of the 11th international conference on information fusion, Cologne, pp 1–8
22. Goubet E, Katz J, Porikli F (2006) Pedestrian tracking using thermal infrared imaging. In: Proceedings of SPIE, vol 62062, pp 62062C–62062C12
23. Hariyono J, Hoang V-D, Jo K-H (2014) Moving object localization using optical flow for pedestrian detection from a moving vehicle. Sci World J 2014:1–8
24. Pathirana PN, Lim AEK, Carminati J, Premaratne M (2007) Simultaneous estimation of optical flow and object state, a modified approach to optical flow calculation. In: Proceedings of IEEE international conference on networking, sensing and control, London, UK, pp 634–638
25. Qi Y, An G (2011) Infrared moving targets detection based on optical flow estimation. In: Proceedings of the international conference on computer science and network technology, China, pp 2452–2455
26. Brehar R, Nedevschi S (2014) Pedestrian detection in infrared images using HOG, LBP, gradient magnitude and intensity feature channels. In: Proceedings of IEEE 17th international conference on intelligent transportation systems, Qingdao, pp 1669–1674
27. Gilmore ET, Ugbome C, Kim C (2011) An IR-based pedestrian detection system implemented with matlab-equipped laptop and low-cost microcontroller. Int J Comput Sci Inf Technol 3(5):79–87
28. Kancharla T, Kharade P, Gindi S, Kutty K, Vaidya VG (2011) Edge based segmentation for pedestrian detection using NIR camera. In: Proceedings of the international conference on image information processing, Himachal Pradesh, pp 1–6
29. Olmeda D, Hilario C, Escalera A, Armingol JM (2008) Pedestrian detection and tracking based on far infrared visual information. In: Proceedings of the 10th international conference on advanced concepts for intelligent vision systems, France, pp 958–969
30. Bertozzi M, Broggi A, Felisa M, Vezzoni G, Del Rose M (2006) Low-level pedestrian detection by means of visible and far infrared tetra-vision. In: Proceedings of the IEEE intelligent vehicles symposium, Tokyo, pp 231–236
31. Torresan H, Turgeon B, Ibarra-Castanedo C, Hebert P, Maldague X (2004) Advanced surveillance systems: combining video and thermal imagery for pedestrian detection. In: Proceedings of SPIE Thermosense XXVI, Vol 5405 of SPIE
32. Tribaldos P, Serrano-Cuerda J, Lopez MT, Fernandez-Caballero A, Lopez-Sastre RJ (2013) People detection in color and infrared video using HOG and linear SVM. In: Proceedings of the 5th international work-conference on the natural and artificial computation in engineering and medical applications, Berlin, Heidelberg, pp 179–189
33. Olmeda D, Escalera A, Armingol JM (2012) Contrast invariant features for human detection in far infrared images. In: Proceedings of IEEE intelligent vehicles symposium (IV), Alcala de Henares, pp 117–122
34. Zin TT, Tin P, Hama H (2011) Pedestrian detection based on hybrid features using near infrared images. Int J Innov Comput Inf Control 7(8):5015–5025
35. Wang X, Han TX, Yan S (2009) An hog-lbp human detector with partial occlusion handling. In: Proceedings of IEEE 12th international conference on computer vision, Kyoto, pp 32–39
36. Schwartz W, Kembhavi A, Harwood D, Davis L (2009) Human detection using partial least squares analysis. In: Proceedings of IEEE international conference on computer vision, Kyoto, pp 24–31

37. Bertozzi M, Broggi A, Caraffi C, Del Rose M, Felisa M, Vezzoni G (2007) Pedestrian detection by means of far-infrared stereo vision. J Comput Vis Image Underst 106(2–3):194–204
38. Dai C, Zheng Y, Li X (2007) Pedestrian detection and tracking in infrared imagery using shape and appearance. J Comput Vis Image Underst 106(2–3):288–299
39. Bertozzi M, Broggi A, Hilario Gomez C, Fedriga RI, Vezzoni G, Del Rose M (2007) Pedestrian detection in far infrared images based on the use of probabilistic templates. In: Proceedings of IEEE symposium on intelligent vehicle, Istanbul, pp 327–332
40. Buch N, Cracknell M, Orwell J, Velastin SA (2009) Vehicle localisation and classification in urban CCTV steams. 16th World Congress and exhibition on intelligent transport systems and services. Stockholm, Sweden, pp 1–8
41. Leibe B, Leonardis A, Schiele B (2004) Combined object categorization and segmentation with an implicit shape model. Workshop on statistical learning in computer vision, Prague, Czech Republic, pp 1732
42. Lin Z, Davis LS (2008) A pose-invariant descriptor for human detection and segmentation. In: European conference on computer vision, Berlin, Heidelberg, pp 423–436
43. Meis M, Oberlander U, Ritter W (2004) Reinforcing the reliability of pedestrian detection in far-infrared sensing. In: Intelligent Vehicles Symposium, pp 779–783
44. Cong DNT, Khoudour L, Achard C, Phothisane P (2009) People re-identification by means of a camera network using a graph-based approach. In: Conference on Machine Vision Applications, Yokohama, Japan, pp 152–155
45. Fei M, Li J, Liu H (2015) Visual tracking based on improved foreground detection and perceptual hashing. Neurocomputing 152:413–428
46. Kushwaha AKS, Srivastava S, Srivastava R (2016) Multi-view human activity recognition based on silhouette and uniform rotation invariant local binary patterns. Multimed Syst J 1–17. doi:10.1007/s00530-016-0505-x
47. Tewary S, Akula A, Ghosh R, Kumar S, Sardana HK (2014) Hybrid multi-resolution detection of moving targets in infrared imagery. Infrared Phys Technol 67:173–183
48. Yin J, Liu L, Li H, Liu Q (2016) The infrared moving object detection and security detection related algorithms based on W4 and frame difference. Infrared Phys Technol 77:302–315
49. Lu N, Wang J, Wu QH, Yang L (2008) An improved motion detection method for real-time surveillance. Int J Comput Sci 35(1):119–128
50. Aptean AD (2011) Contributions to the information fusion: application to obstacle recognition in visible and infrared images, PhD thesis, National Institute of Applied Sciences of Rouen, France and Technical University of Cluj-Napoca, Romnia
51. Wang J, Liang J, Hu H, Li Y, Feng B (2007) Performance evaluation of infrared and visible image fusion algorithms for face recognition. In: Proceedings of the international conference on intelligent systems and knowledge, Engineering, pp 1–8
52. Davis JW, Sharma V (2007) Background-subtraction using contour-based fusion of thermal and visible imagery. Comput Vis Image Underst 106(2–3):162–182
53. Xiao G, Wei K, Jing ZL (2008) Improved dynamic image fusion scheme for infrared and visible sequence based on image fusion system. In: Proceedings of the 11th international conference on information fusion, Cologne, pp 1745–1750
54. Lewis JJ, O'Callaghan RJ, Nikolov SG, Bull DR, Canagarajah CN (2007) Pixel and region based image fusion with complex wavelets. J Inf Fusion 8(2):119–130
55. Lanir Y (2005) Comparing multispectral image fusion methods for a target detection task. Thesis submitted in partial fulfillment of the requirements for the M.Sc Degree
56. Yang B, Zhong-liang J, Hai-tao Z (2010) Review of pixel-level image fusion. J Shanghai Jiaotong Univ 15(1):6–12
57. Pop S (2008) Modle de fusion et diffusion par quations aux drives partielles: application la sismique azimutale, PhD thesis, Bordeaux I University, France
58. Blum RS, Xue Z, Zhang Z (2006) An overview of image fusion. In: Liu Z, Blum RS (eds) Multi-sensor image fusion and its applications. CRC Press, pp 1–36
59. Li M, Dong Y (2013) Review on technology of pixel-level image fusion. In: Proceedings of the international conference on measurement, information and control, Harbin, pp 341–344
60. Jagalingam P, Hegde AV (2014) Pixel level image fusion a review on various techniques. In: Proceedings of the 3rd world conference on applied sciences, engineering and technology, Kathmandu, Nepal, pp 521–528
61. Yang B, Jing Z-liang, Zhao H-tao (2010) Review of pixel-level image fusion. J Shanghai Jiaotong Univ Sci 15(1):6–12
62. Amro I, Mateos J, Vega M, Molina R, Katsaggelos AK (2011) A survey of classical methods and new trends in pansharping of multispectral images. EURASIP J Adv Signal Process 2011(1):1–22
63. Singh S, Gyaourova A, Bebis G, Pavlidis I (2004) Infrared and visible image fusion for face recognition. Proc SPIE 5404:585–596
64. Desale RP, Verma SV (2013) Study and analysis of PCA, DCT & DWT based image fusion techniques. In: Proceedings of IEEE international conference on signal processing image processing and pattern recognition, Coimbatore, pp 66–69
65. Haghighat MBA, Aghagolzadeh A, Seyedarabi H (2010) Real-time fusion of multi-focus images for visual sensor networks. In: Proceedings of the 6th IEEE Iranian conference on machine vision and image processing, Isfahan, pp 1–6
66. Sadjadi F (2005) Comparative image fusion analysais. In: IEEE Computer Society Conference on computer vision and pattern recognition—workshops, vol 8, No (8), pp 25–25
67. Hammami M, Jarraya SK, Ben-Abdallah H (2013) On line background modeling for moving object segmentation in dynamic scene. Multimed Tools Appl 63(3):899–926
68. Zhang E, Zhang Y (2009) F-Measure. In: Liu L, Özsu MT (eds) Encyclopedia of database systems. Springer, New York
69. Kim K, Chalidabhongse T, Harwood D, Davis L (2005) Real-time foreground-background segmentation using codebook model. Real Time Imaging J 11(3):167256
70. Parag T (2014) Enforcing label and intensity consistency for IR target detection. In CoRR abs, pp 1–21
71. Mouats T, Aouf N (2014) Fusion of thermal and visible images for day/night moving objects detection. Sensor Signal Processing for Defence (SSPD), Edinburgh, pp 1–5
72. Mangale S, Khambete M (2016) Camouflaged target detection and tracking using thermal infrared and visible spectrum imaging, Intelligent Systems Technologies and Applications, pp 193–207

Emna Fendri obtained the engineering degree in computing of the National Ecole of The sciences of computing, Tunisia (ENSI). She received a Ph.D in computer science from the University of Sfax in Tunisia (FSEGS) in 2010. She is currently assistant professor in the Computer Science Department at the Faculty of Science Sfax, Tunisia. She is a researcher in the MIRACL Laboratory (Multimedia, InfoRmation systems and Advanced Computing Laboratory). Her current research interests include multimedia indexing and retrieval, video surveillance and multimedia mining.

Rania Rebai Boukhriss graduated with a master's thesis in computer science from the University of Sfax in Tunisia. She is a researcher in the MIRACL Laboratory (Multimedia, InfoRmation systems and Advanced Computing Laboratory). Currently, she is preparing her Ph.D. at the University of Sfax in Tunisia (FSEGS). Her research interests include computer vision, video and image processing.

Mohamed Hammami received a Ph.D. in computer science from EcoleCentrale at the Lyon Research Center for Images and Intelligent Information Systems (LIRIS) associated with the French research institution CNRS as UMR5205. He is currently associate professor in the Computer Science Department at the Faculty of Science Sfax, Tunisia. He is a researcher in the MIRACL Laboratory (Multimedia, InfoRmation systems and Advanced Computing Laboratory). His current research interests include data mining and knowledge discovery in images and video, multimedia indexing and retrieval, face detection and recognition, and Web site filtering. He was a staff member in RNTL-Muse project. He has served on technical conference committees and as reviewer in many international conferences.