Advanced Surveillance Systems:
Combining Video and Thermal Imagery for Pedestrian Detection

H. Torresan, B. Turgeon, C. Ibarra-Castanedo, P. Hébert, X. Maldague

Electrical and Computing Engineering Dept.,
Université Laval, Québec (Québec), G1K 7P4, Canada[1]

Key words: surveillance, security, infrared, video imagery, image processing

Abstract

In the current context of increased surveillance and security, more sophisticated surveillance systems are needed. One idea relies on the use of pairs of video (visible spectrum) and thermal infrared (IR) cameras located around premises of interest. To automate the system, a dedicated image processing approach is required, which is described in the paper. The first step in the proposed study is to collect a database of known scenarios both indoor and outdoor with a few pedestrians. These image sequences (video and TIR) are synchronized, geometrically corrected and temperature calibrated. The next step is to develop a segmentation strategy to extract the regions of interest (ROI) corresponding to pedestrians in the images. The retained strategy exploits the motion in the sequences. Next, the ROIs are grouped from image to image separately for both video and TIR sequences before a fusion algorithm proceeds to track and detect humans. This insures a more robust performance. Finally, specific criteria of size and temperature relevant to humans are introduced as well. Results are presented for a few typical situations.

1. Introduction

In the current context of increased surveillance and security, the necessity has emerged for more sophisticated surveillance systems, for instances around buildings. One idea that is promising relies on the use of pairs of video (visible spectrum) and thermal infrared (IR) cameras distributed around premises of interest. Since it is not practical to have humans observing the resulting images in real-time, it is proposed to add an “intelligent fusion/detection step” to the system so that human observers are involved only in case “abnormal situations” occur.

Combining visible and thermal infrared images (TIR) is advantageous since visible images are much affected by lighting conditions while TIR images provide enhanced contrast between human bodies and their environment. However in outdoor conditions, it was noticed that TIR images are somewhat sensitive to wind and temperature changes. Nevertheless, these limitations for both modalities are independent and usually do not occur simultaneously. An intelligent fusion of the information provided by both sensors reduces detection errors, thereby increasing the performance of tracking and the robustness of the surveillance system.

A literature search reveals a few interesting papers on the exploitation of near - infrared information to track humans [1-3]. These papers generally deal only with the face of observed people [1-2] and a few are concerned with the whole body [3]. However, when looking to the efforts in the visible part of the spectrum for the same task, many papers are available such as [4,5]. Surprisingly, the idea to couple visible and thermal infrared is not yet seen as a popular research field for this application. One reason explaining this is probably due to the still high cost of the thermal infrared cameras (~ 10 k$) versus their visible counter parts (~ 1-2 k$ for quality products). Moreover outdoor scenarios are obviously more challenging to visible imagery due to shadows, light reflections, levels of darkness and luminosity. However, with Planck’s law [6] in mind, it is clear that thermal infrared in bands 2-5 μm and especially 8-12 μm offers better “light immunity” from the sun whose emitted radiation peaks at 0.5 μm right in the middle of the visible spectrum (0.4-0.8 μm). Nevertheless, moving leaves and grass, cooling winds, moving shadows with clouds, reflecting snow, etc., are challenging for TIR imagery.

1. maldagx@gel.ulaval.ca

As stated in the title, the goal of the present work is to track pedestrians. This means: *walking and standing humans*, excluding all “upside-down” - ! - or sitting people. No constraints are put on clothes (we tested our system in all four seasons with subjects wearing from light clothes to heavy coats). The last restriction concerns the number of pedestrians present in the scene. If too much pedestrians are present, the number of moving objects in the scene becomes too large compared to the background and many pedestrians are simply represented by a blob. The fact that our system exploits the two modalities and works with both indoor and outdoor scenes is not common in the literature.

In the paper, the acquisition system is first briefly presented. Next, the processing algorithm for pedestrian extraction is presented. The paper ends with a presentation of a few results.

2. ACQUISITION SYSTEM

The acquisition system is composed of a ’mobile platform’ on which the two cameras with their own computer are mounted. The system was moved around inside and outside our pavilion so that different scenarios were recorded. The video camera is a 640 x 480 pixels, Pulnix TMC-6700CL color CCD connected to a Meteor II Camera Link Matrox frame grabber while the IR is a 320 x 256 pixels, InSb SBF connected to a Genesis LC Matrox frame grabber. Because of the non-versatile commercial acquisition software of the IR-camera, two computers were used. Infrared images undergo a series of pre-processing steps to correct them for vignetting, fixed pattern noise, dead pixel and finally temperature calibration [6].

Due to their different resolution (IR: 320 x 256 pixels and visible: 640 x 480 pixels) and non perfect colinearity alignment, visible and TIR images cannot be compared “pixels by pixels.” We have thus developed a “region-based” rather than “pixel-based” approach; all proceeds in a relative rather than absolute manner. A geometric calibration is done on both separate cameras to obtain intrinsic parameters [7]. The intrinsic parameters provide a scale factor in x and y, the focal lens, and the radial distortion.

Moreover, in the experiments, time synchronization is also performed manually through the observation of both scenes with our dedicated program. It is to be noted that only the beginning of the sequence needs to be synchronized because the acquisition frame-rate is known and accurate for both cameras.

3. IMAGE PROCESSING

The main part of the work concerns image processing. The goal here is to extract pedestrian(s) from sequences. In this section, it is assumed video and IR sequences are acquired and pre-processed as described in the previous section. An important hypothesis is that cameras do not move during the recording of one given sequence (which is the case in most surveillance systems). Figure 1 presents the overall image processing algorithm. After image acquisition, the second step consists in extracting moving regions with a background subtraction algorithm. The next step deals with the tracking of the moving blob. Each blob is analysed and eventually combined to create or update an object. In the fourth step, tracking of the object is implemented for each modality. The next step involves the correspondence between objects of the visible and IR images to identify fusion hypotheses between objects. These hypotheses can be fed back to the object tracking module for each modalities. The last two steps correspond to detecting occlusion between fused objects and using the statistics of the object to determine whether the object is a pedestrian or not. It is worth noting that in the proposed architecture, sensor fusion is applied at the level where an object representation is available. This modular architecture differs from an approach where fusion would occur at each step. It builds on the processing pipeline of single modalities.

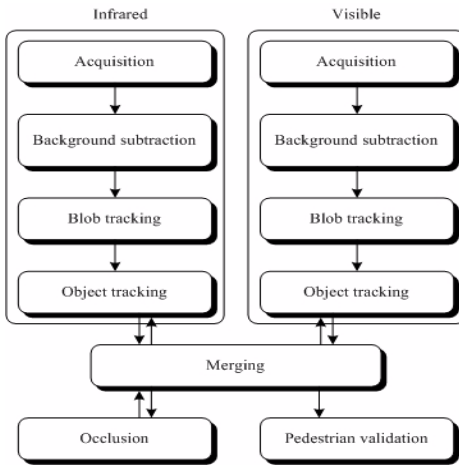

Figure 1: Image processing architecture

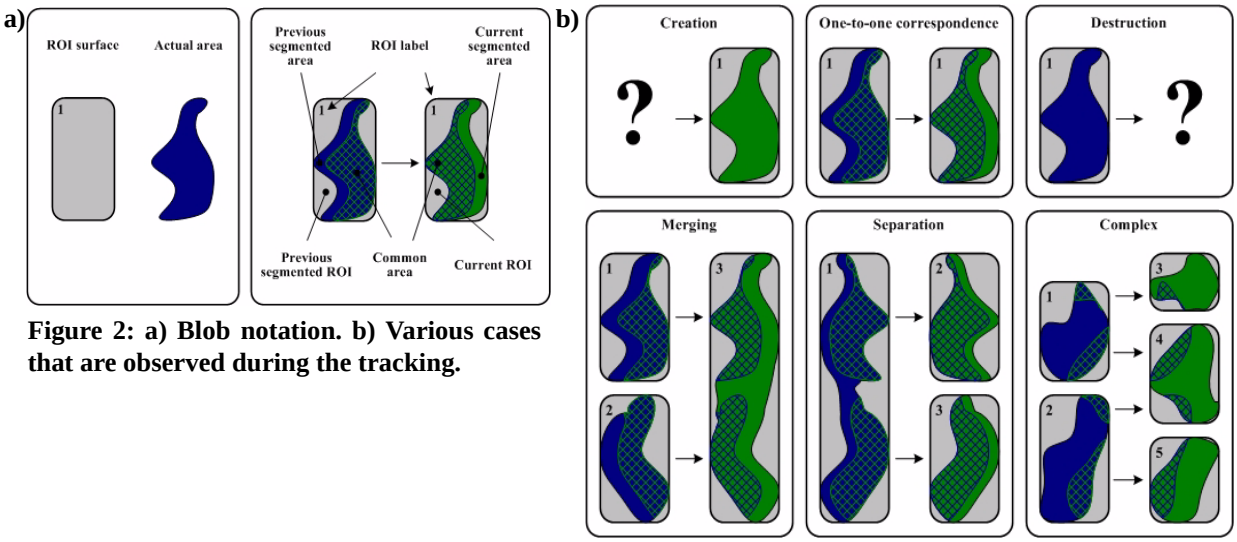

Figure 2: a) Blob notation. b) Various cases that are observed during the tracking.

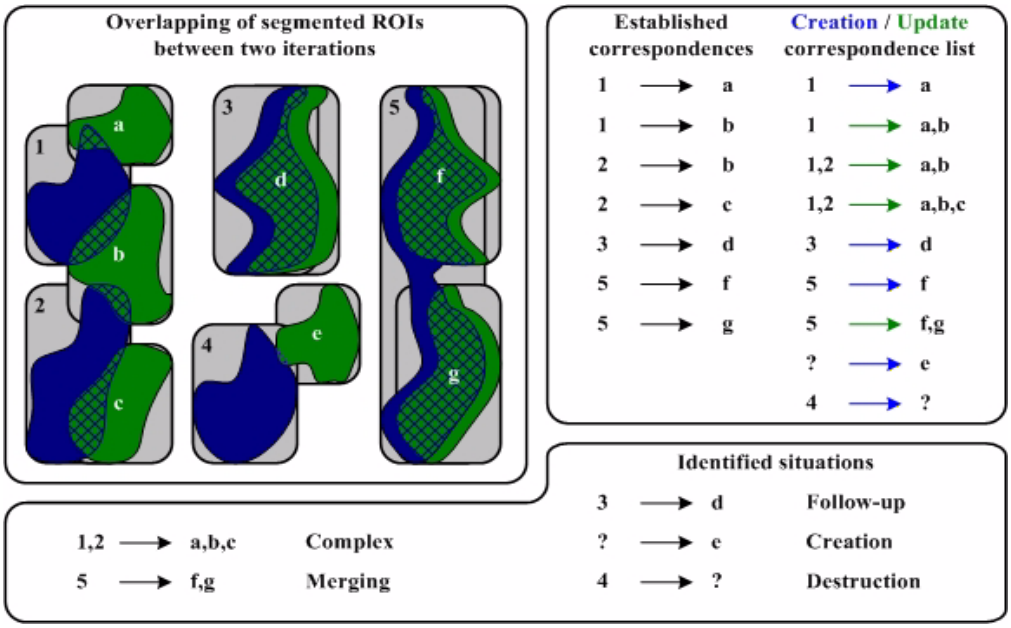

Figure 3: Examples of list correspondences to follow blobs.

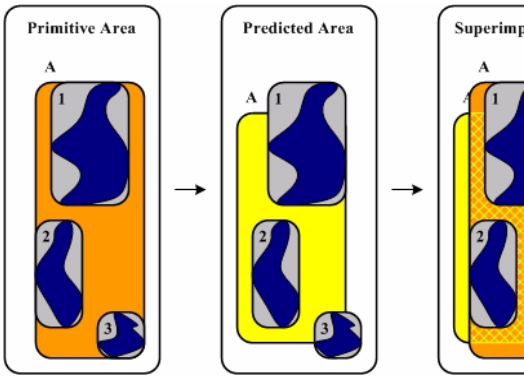

Figure 4: Object primitive (orange) and predicted areas (yellow).

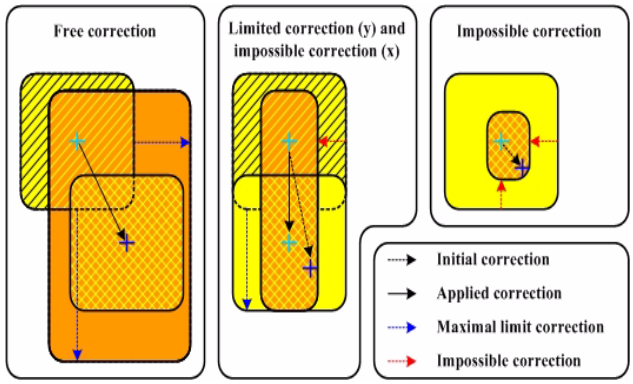

Figure 5: Representation of the updated state for the predicted area.

3.1 Background subtraction

For the visible and IR sequences, an adaptive background subtraction algorithm [8] is performed that consists in accumulating statistics (mean and standard deviation) for each pixel in the image and then comparing these statistics with the pixel values in the newly acquired image. If the vector value (RGB for visible images and temperature for IR images) is too different according to statistics, one considers a pixel as a foreground pixel. After all pixels have been classified as foreground or background pixels, statistics are updated using the last image. This adaptive algorithm is more robust to low light levels or temperature change during the day. For short sequences (less than 2 min.), 30 frames are typically required at the beginning of the sequence without any pedestrian to initialize statistics of the background.

The blobs in the image are obtained by using an eight-neighborhood connected component algorithm on the foreground pixels. These blobs are used in the first level tracking algorithm. For speeding up the tracking process, the region of interest (ROI) is rectangular and corresponds to the bounding box of the blob (See Figure 2a).

3.2 Two-level tracking

Tracking is performed at two levels. While the first level of the tracking algorithm consists in following the blobs in an image sequence, the second level builds on the first and tracks a combination of one or more blobs, i.e. objects. To do this, some feature parameters between blobs at time 't' and 't-1'are first introduced.

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

$$O_{max}(a,b) = Maximum(CS(a,b) / AROI(a), CS(a,b) / AROI(b)) , \quad (1)$$

$$O_{min}(a,b) = Minimum(CS(a,b) / AROI(a), CS(a,b) / AROI(b)) , \quad (2)$$

where *AROI(i)* is the area of the i^{th} blob's ROI and *CS(a,b)* is the intersection area between the two ROI.

"Similarity", *S(a,b)*, is defined as:

$$S(a,b) = 1 - [Abs(A_a - A_b) / Maximum(A_a, A_b)], \quad (3)$$

where A_i is the actual area of the i^{th} blob (see Figure 2a).

"Resemblance", *R(a,b)*, between two ROI a and b is defined as:

$$R(a,b) = [O_{min}(a,b) \times S(a,b)]. \quad (4)$$

During tracking, the maximum overlapping factor (Equation 1) is used to follow-up blobs between two frames of a sequence. Figure 2b depicts all possible cases of blob tracking. When a one-to-one correspondence is obtained, the same label is given to the blob of the new frame. When a complex case is obtained, a more accurate analysis must be carried out so as to reduce blobs of the complex case to a simpler case such as in Figure 2b (one-to-one correspondence, merging, separation, creation or destruction). An algorithm computing the resemblance factor between all of the blobs is used to simplify the complex case. The resemblance factor eliminates much more of the correspondence between the blobs, since it is based on the minimum overlapping and similarity factor. Following this, specific parameters, like the speed and the confidence, are computed for the blob. The confidence (C) is a feature that gives the persistence of a blob over time and is described by the following equation:

$$C(a) = \left(\sum_{b=0}^{n} R(a,b) \times C(b)\right) + 1 , \quad (5)$$

where *a* is for the new blob, *b* the preceding blob and *n* the number of preceding blobs that are greater than one in a merging or a complex case.

As seen from Equation 5, the confidence on matching from *t-1* to *t* increases if the blob has been tracked for a long time and the resemblance from two time steps is large. Obviously, confidence is zero at $t = 0$. Figure 3 gives an example of blob tracking when the complex simplification algorithm is used: the complex case (1,2 -> a,b,c) can be reduced to (1->a,b) and (2->c). Moreover to make sure no hypothesis is lost, all blobs are followed in the sequences, not only those "believed to belong" to pedestrians.

Before we can detect a pedestrian, all blobs must be grouped together to create objects. An object can be made up from several blobs. An object is initially created from an isolated blob or many closer blobs. The object has a primitive area (PrA) that is generated from the ROI of each blob of the object and a predicted area (PdA) that evolves (size, position and speed) differently from the primitive area (See, Figure 4). As time passes, an object at time *t* inherits the blob(s) making up the object at time t-1. A blob that appears in a predicted area is also added to the object. A blob that has an opposite movement from the object can be removed from the object. The object also has confidence that is computed as the average of the confidence of individual blobs comprising the object. The predicted area is an important feature that gives the real size, position and speed of the object. If no blob has been detected for a long period of time, the object vanishes. A non-moving object where blobs appear and disappear in time may be labelled as a noisy object such as a moving leaf in a tree.

The adjustment of the position of the predicted area is first estimated using the mean speed of the predictive area of the last fifteen frames. Then, an algorithm corrects this estimated position by using the center of mass of the primitive area. This correction is limited by the difference between the position of the boundary of the predictive and primitive areas. Three examples illustrating the adjustment of the position of the predicted area are presented in Figure 5. Finally, the mean speed is updated with the new position value.

3.3 Merging

The merging algorithm is driven by three goals. The first one consists in establishing a correspondence between the objects detected in the visible and the IR images. For each pair of objects, the identification of the best object detected (in visible or IR images) describes our second goal. The object with the best detection will be called *master* and the second one *slave*. The confidence is used as a criterion for better detection and is computed for all the objects of each frame in the sequence. In this manner the identification of the master and the slave will change rapidly for an object when fast light illumination or temperature variation are present. Our last goal consists in using the information of the *master* object to help in tracking the *slave* one. The merging process is done independently for each pair of objects. For example, if at time *t*, three objects can be detected in the visible and infrared images, two objects can be *master* in the infrared image, and one object can be a *master* in the visible image.

The merging algorithm can modify the position and the size of the predicted area computed during the *second level of tracking*. But this only occurs when a great difference between the primitive area of the master object and the slave object is detected. In this case we enter in the "enslavement" mode where the *master* predicted area controls the *slave* predicted area. For example, if a pedestrian has a green T-shirt and walks in front of a green hedge, this person's trunk will tend to disappear and the *slave* object will be put in the enslavement mode. The IR object will maintain a good detection and will help in tracking the pedestrian in the visible image because the body temperature is higher than that green hedge temperature.

In the case where two objects disappear, objects will stay present in the system and the position of the predictive area is assessed using the mean speed of the predictive area in the last frame. For example, if a pedestrian passes behind a tree, the objects will disappear in both images. If the pedestrian maintains his speed and direction, the object will be recovered when it appears on the other side of the tree. But if the pedestrian stops behind the tree and returns to the same side, the algorithm will create a new object.

3.4 Occlusion detection

Since the system tracks many blobs and objects, it supports tracking of many pedestrians and deals with occlusions as well. For an occlusion to occur we first need to have two objects that have been successfully associated and tracked

by the merging algorithm. So, a merging object cannot be in occlusion with another object detected by only one of the sensors and probably representing noise. To be merged, objects have to be at a minimal distance from other merged objects. So, two different parts of one pedestrian cannot be in occlusion and will probably be combined later with the object already merged (since they make up that pedestrian). An occlusion occurs when blobs of two or more objects detected by the two sensors are merged during the tracking. For example, during the tracking of the blobs, a blob A belonging to an object 1 is combined with a blob B belonging to an object 2. A case of merging (Figure 2b) between the blob A and the blob B is then identified and an occlusion occurs, since the blobs belong to two different objects detected by the two sensors. When objects are in an occlusion state, the position of the predictive area is predicted for each object. When blob separation is then detected, a verification is performed to determine whether each blob can correspond to each object. If this is the case the occlusion is solved. An example of occlusion is presented in Figure 8.

3.5 Pedestrian validation

Three criteria are examined in order to determine whether or not a merged object is a human: the aspect ratio (length over width), temperature and also step frequency. When a person walks, a certain frequency can be observed in the manner in which she/he moves her/his arms and legs. It is the width of the predicted area which is used to compute the step frequency with a FFT. This frequency helps to eliminate objects such as cars whose size does not vary periodically. At each frame of the sequence, these three criteria are compared to a lower and upper bounds that represent a pedestrian. If all of the criteria are respected in infrared or visible objects, a vote is taken. The number of votes is computed and compared to the number of frames to give a percentage of pedestrian detection. For example, if the three criteria are met over 30 frames in a 70 frame acquisition sequence, the detection of the pedestrian will be at 42%. It is also possible to give different comparison ranges for different types of objects like cars and trucks where for example the frequency will be zero.

4. RESULTS

The algorithm described in the previous sections was tested on several sequences. Figures 6, 7, 8 illustrate various cases. It is obviously not possible to render the dynamics of these sequences in a paper and thus, some interesting situations were selected. In Figure 6, an outdoor situation of one pedestrian walking near a wall is presented and shows that the IR image can be helpful in removing shadows from the visible image. In Figure 7 two outdoor pedestrians are shown where the blobs of one pedestrian are not well detected in both IR and visible images. The merging algorithm improved detection for the predicted area of this pedestrian. Finally Figure 8 presents some snap shots of a sequence of two crossing pedestrians to show the robustness of the occlusion algorithm.

The results shown illustrate the robustness of the system. In a future configuration, we want to have both IR and visible cameras linked to a common acquisition system on a single computer. This will enable real-time work on the sequences instead of off-line work as of now. We want also to work with more than one “node.” A node consists in one visible-IR camera pair. In a multi-node system, interaction between several nodes would allow tracking a given individual from different points of view.

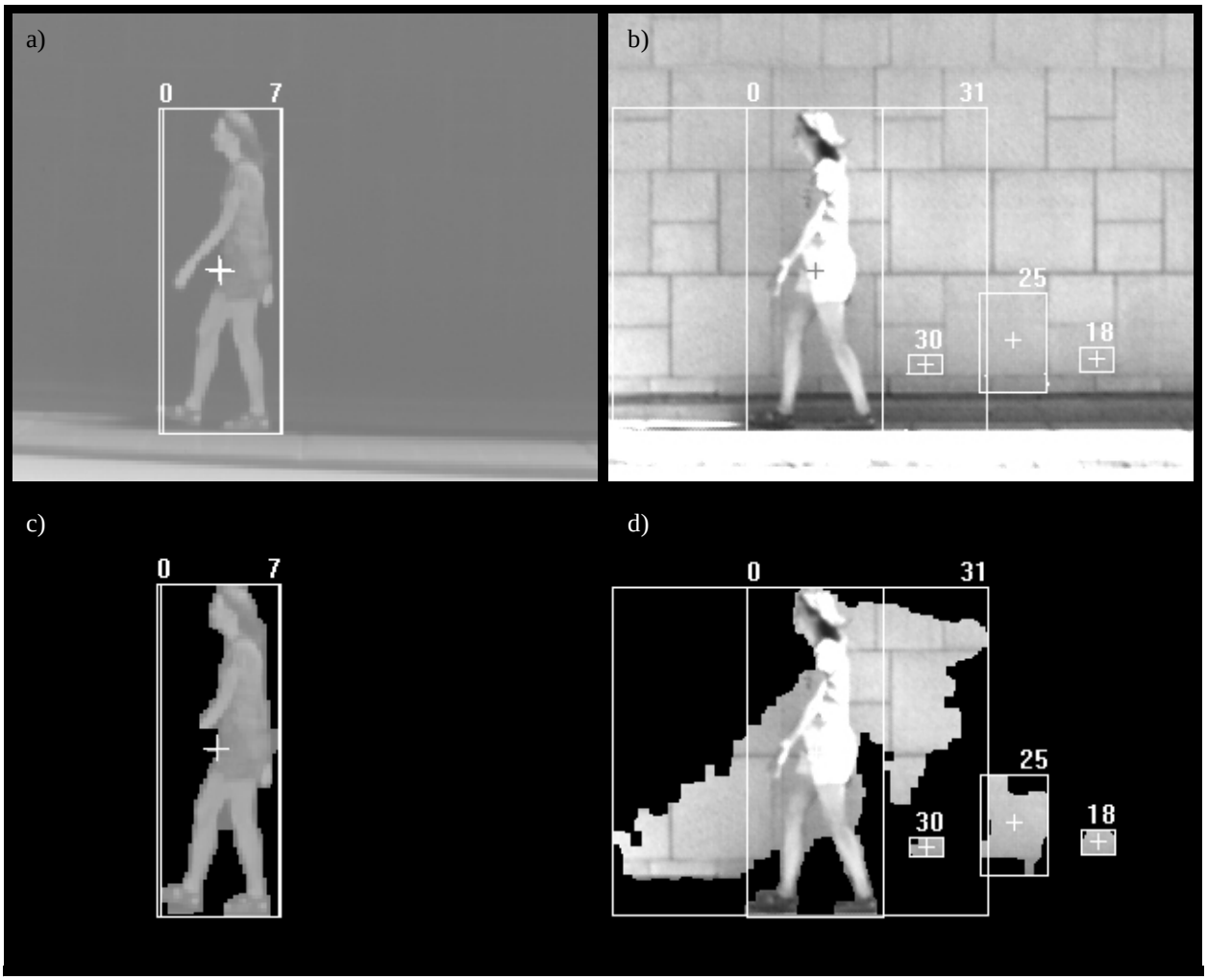

Figure 6: Outdoor scene illustrating pedestrian extraction. a,b) Original IR and visible images. c,d) Representation of the blob detected for both IR and visible images. Note that the blob in the visible image also includes the shadow of the person. But the predicted region (labelled with zero in the upper-left corner) is the same in all pictures. The ROI of blobs (labelled 18, 25, 30 and 31 to the upper-right corner) are very different from the predicted regions in the visible image. The ROI of blob (labelled 7 in the upper-right corner) is similar to the predicted region labelled 0.

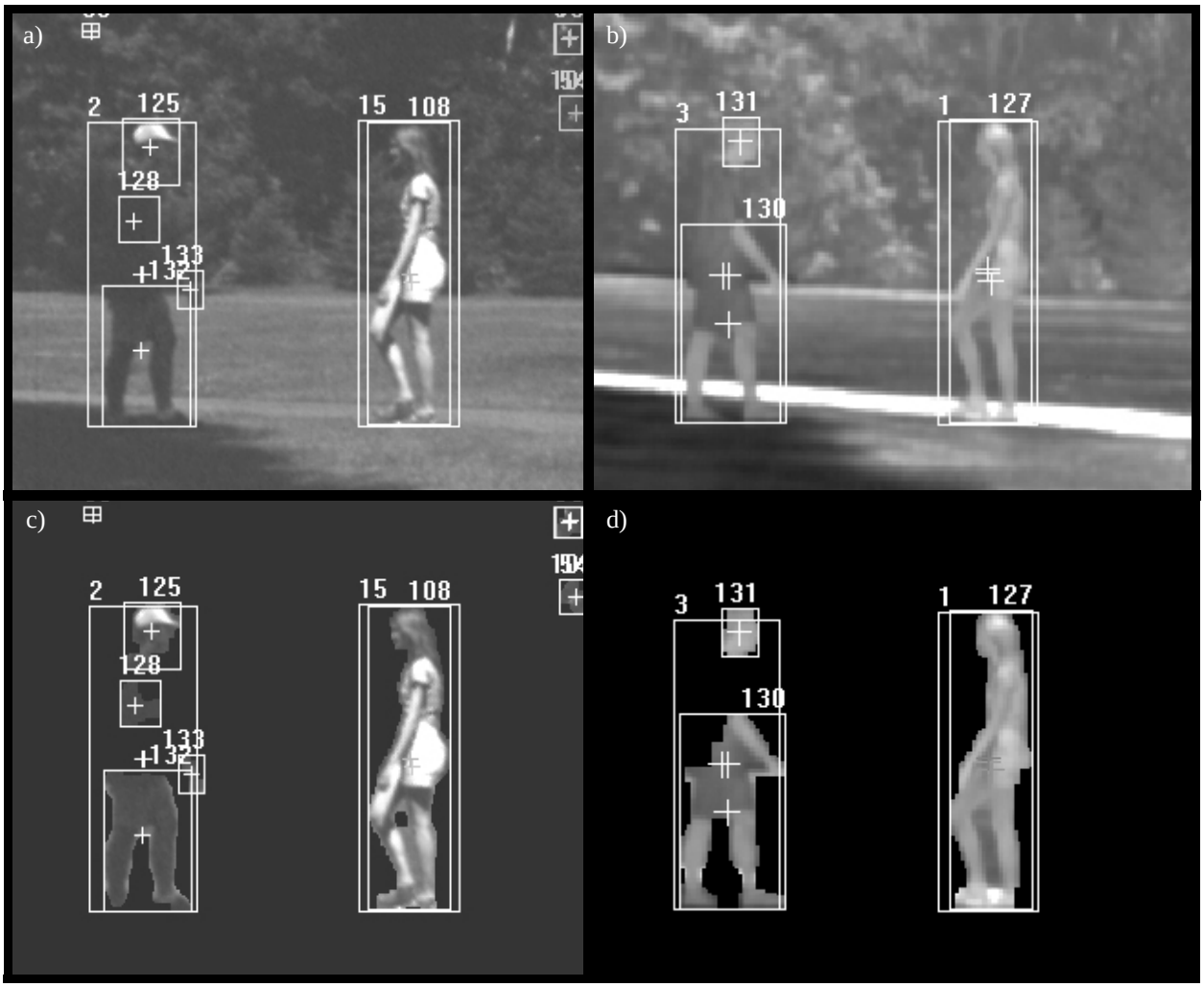

Figure 7: Outdoor scene showing a two pedestrian extraction. a,b) Original IR and visible images. c,d) Representation of the blob detected for both IR and visible images. The rectangles with the label in the upper-left corner give the predicted area numbers (1,2,3,15) and the rectangles with the label in the upper-right corner give the blob number. We can see that the left pedestrian was not completely detected by the background subtraction algorithm in both IR and visible images. Meanwhile, the predicted area is well estimated for the two pedestrians.

Figure 8: Outdoor scene showing two pedestrians tracked in infrared (upper sequence) and visible (lower sequence) for six snap shots. The labels 1 and 3 represent the two predicted area of the pedestrian in the infrared images and the labels 2 and 15 represent the two predicted areas of the pedestrian in the visible images. The other labels represent the blobs detected by the background subtraction algorithm.

5. CONCLUSION

In this paper, a system to automatically track pedestrians in simultaneous visible and IR sequences was presented. The system features: outdoor operation, multiple (up to three) pedestrians at a time, no need for exact matching of visible and IR cameras since the system is area-based rather than pixel-based. Image processing was discussed and some typical results were presented for three different sequences.

6. ACKNOWLEDGEMENTS

The supports of the Natural Sciences and Engineering Research Council of Canada and of the *Fonds FQRNT* of the Province of Québec are acknowledged.

REFERENCES

1. C. K. Eveland, D. A. Socolinsky, L. B. Wolff, "Tracking human faces in infrared video," *Image and Vision Computing*, **21** (2003): 579-590.
2. J. Dowdall, I. Pavlidis, "Face detection in the near-IR spectrum," Image and Vision Computing, **21** (2003): 565-578.
3. M. Bertozzi, A. Broggi, P. Grisleri, T. Graf, M. Meinecke, "Pedestrian Detection in Infrared Images," In *Proc. IEEE Intelligent Vehicles Symposium 2003*, 662-667, Columbus (USA), June 2003.
4. O.Masoud, N. Papanikolopoulos, "A method for human action recognition," *Image and Vision Computing*, **21** (2003): 729-743.
5. I. Haritaoglu, D.Harwood, L. S. Davis, "W4: Real-Time Surveillance of People and Their Activities," *IEEE Transactions On Pattern Analysis And Machine Intelligence*, **22**[8]: 809-830, 2000.
6. X. Maldague, *Theory and Practice of Infrared Technology for Non Destructive Testing*, John-Wiley & Sons, 684 p., 2001.
7. Heikkilä J., Silvén O., "A Four-step Camera Calibration Procedure with Implicit Image Correction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **22**[10]: 1066-1077, 2000. [see also Matlab source code at: www.vision.caltech.edu/bouguetj/calib_doc/]
8. A. Lemieux, M. Parizeau, "Flexible multi-classifier for face recognition systems" *Vision Interface*, S1.4, 2003 (http://kopernik.eos.uoguelph.ca/~zelek/vi2003/).