Real-Time Posture Analysis in a Crowd using Thermal Imaging

Quoc-Cuong Pham, Laetitia Gond, Julien Begard, Nicolas Allezard, Patrick Sayd
CEA, LIST,
Boîte Courrier 65, Gif-sur-Yvette, F-91191 France
quoc-cuong.pham@cea.fr

Abstract

This article describes a video-surveillance system developed within the ISCAPS project. Thermal imaging provides a robust solution to visibility change (illumination, smoke) and is a relevant technology for discriminating humans in complex scenes. In this article, we demonstrate its efficiency for posture analysis in dense groups of people. The objective is to automatically detect several persons lying down in a very crowded area. The presented method is based on the detection and segmentation of individuals within groups of people using a combination of several weak classifiers. The classification of extracted silhouettes enables to detect abnormal situations. This approach was successfully applied to the detection of terrorist gas attacks on railway platform and experimentally validated in the project. Some of the results are presented here.

1. Introduction

The general objective of the ISCAPS[1] project is to reduce the risks or the consequence of malicious events by providing efficient, real-time, user-friendly, highly automated surveillance of crowded area. This article describes one of the surveillance systems developed within ISCAPS to cope with a realistic scenario specified by an end-user involved in the project (a national railway company). In this scenario, the scene takes place on a railway platform and the system must automatically detect gas attacks from the behaviour analysis of people, with the shortest delay. Indeed, evidence indicates that early detection of the use of chemical agents is of paramount importance to reduce the number of casualties [13]. Two functionalities are important. First, if smoke, caused by a gas attack or a fire, fills the area, a panic will probably occur. Most of people are going to leave the area, but some persons may be blocked in a highly dangerous environment. Second, short of gas detectors, serious incidents can be detected by observing the reaction of people if many people stagger, with heavy coughing, or fall down. The scenario combines seeing into darkness or smoke and image interpretation of people becoming unwell, both technically challenging. A un-cooled infrared sensor (micro-bolometer technology, $8 - 12\mu m$) is used to bring robustness to hard visibility conditions (figure 1). In our scenario, we have to detect persons present on the platform filled by smoke and also persons falling down. Due to their expensive cost and their reduced life expectancy, the use of far infrared sensors was limited to military applications in order to detect and track people or vehicles. Thanks to the new generation of un-cooled infrared sensors cheaper and more robust than cooled ones, a new field of use is open today to this technology. The high performance of these sensors for bad visibility context and for people detection (thanks to natural human infrared emission) is promising for their spread to surveillance applications as site monitoring [3] or driving assistance [12, 15]. In ISCAPS, the robustness of infrared sensors with respect to smoke was the main criterion for the use of this technology. However, infrared images are monochromatic and the object texture is quite poor regarding to visible spectrum images.

Figure 1. Influence of smoke in both color and thermal images. First row: no smoke, second row: area filled with smoke.

The scene analysis is very challenging because of the

[1] http://www.iscaps.net

1-4244-1180-7/07/$25.00 ©2007 IEEE

density of people in front of the camera and also the possible presence of luggage on the platform.

Regardless of the imaging technology, most of conventionnal visual surveillance approaches focus on the detection and tracking of individuals with a reduced overlap. The complexity of crowded scenes (number of people, occlusions, variability of posture) requires specific techniques. In [4], a method is proposed to estimate crowd density and motion based on optical flow and edge extraction. But, the method is not reliable in case of very overlapped people. In [10], authors aim to manage globally the tracking of a group and the estimation of its density. In [11], the density of crowd is estimated by estimation of the fractal dimension of edges. [1] presents an event detector for emergencies in crowds, based on optical flow statistics extracted from the crowd video data. However, these approaches are not well-designed to detect abnormal behaviour of some individuals in the crowd. In our application, a mandatory step consists in the extraction of individuals in the crowd. There have been some approaches to address this problem. In [17], a Bayesian model based segmentation algorithm is proposed using shape models in order to count people. But, this method is prohibitively slow for large crowds. In [14], a crowd detection algorithm based on spatio-temporal analysis of a sequence is presented. The segmentation of moving regions is combined with classification of pedestrian, crowd and vehicles. This approach is dedicated to counting of people more than individuals extraction and cannot analyze fixed people.

Our approach consists in infrared image processing to extract individuals from dense groups and propose an innovative solution to detect people lying down.

2. Overview of the proposed method

Since the IR camera is static, it is clearly advantageous to model the background and to segment foreground objects in a first preprocessing module. The background is learned using an adaptive statistical approach. Because of the variability of clothes and posture, and the potentially high local density of people in the image, head shape and appearance appear to be the most stable visual features over individuals. In the second stage of our algorithm, hypotheses of individuals are generated by a head detector which combines three complementary techniques: i) a local peak detection in the foreground map, ii) an elliptical shape detection, and iii) a head-shoulder pattern detection. The head hypotheses are then classified in two groups. Detected heads above a given height threshold are used to initialize models of human standing. The parameters of the human standing models are refined in a segmentation stage to perform a more accurate localization of people standing in the image. The segmentation step enables to detect blobs in the foreground map located near the ground, and label them as *people lying down* or other object. Below the height threshold, the detected heads form hypotheses of people lying down and directly input the threat detection module which estimates the risk in a probabilistic framework. The workflow of our method is presented in figure 2.

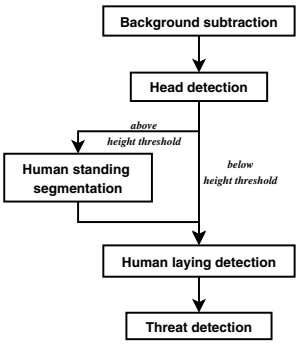

Figure 2. Overview of the algorithm.

3. Human segmentation in IR images

3.1. Background modelling

In thermal images, bodies can be more easily visually discriminated from the background than in color images. However, there exist other specific issues related to the addressed technology. For example, intensities may vary from an individual to another and largely depend upon the environmental conditions [5]. Other effects, like the changing thermal polarity of objects or the nonhomogeneity of bodies, add complexity to the segmentation of persons. Although less sensitive to the lighting conditions than color imaging, infrared emission due to the sun illumination on objects are visible in thermal images, as shown in figure 3.

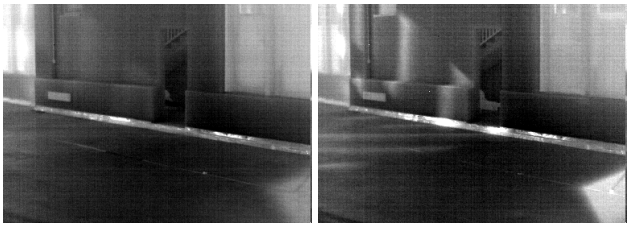

Figure 3. Background scene with different illumination conditions.

Sequential Kernel Density Approximation (SKDA) has proven to be effective to model background [7] or appearance for tracking algorithms [8]. The robustness of this method comes from its ability to encode multiple modes, to adapt to smooth variations over time by integrating weighted new samples and progressively forgetting oldest samples. The SKDA algorithm gives a compact representation because modes which are close to each other can be

merged with a mean-shift mode finding procedure, in a linear time complexity [8].
A specificity of our IR sensor is that a small shift of background intensities can occur, depending on the body sources present in the field of view of the camera. This undesirable effect can lead to poor performance of the background subtraction process. To overcome this problem, we propose to perform a two-pass subtraction: after the first background subtraction, we compute the offset between the mean of the SKDA model modes and the mean of pixels labelled as background, and apply this offset in the second background subtraction. In some cases, the obtained foreground map is substantially improved after the second pass.

3.2. Configuration of the acquisition system

The scenario of interest takes place on a railway platform where people are waiting for a train. For our sensor, only a $25\,mm$ focal length was available. Consequently, to cover the largest area and taking into account the site constraints, the sensor is located at $10\,m$ in front of the platform. In this configuration, the depth of the platform is small with respect to its width. We approximate it as a vertical plane in the 3-D world limited at the bottom by a line on the ground at height $z = 0$ and at the top a maximal line at $z = 2\,m$ (see figure 4). A second approximation consists in defining a direct mapping between the 3-D plane and the region of interest in 2-D in the image, *e.g.* for a given horizontal position in the image the 2-D distance in pixels between the two lines corresponds to a real height of $2\,m$.

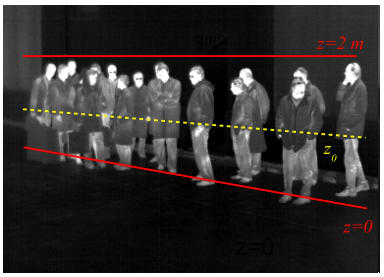

Figure 4. The platform is approximated by 3-D plane from the sensor point of view. The area of interest is two lines at $z = 0$ and $z = 2\,m$. The dot line represents a height threshold z_0 that will be used for detecting people lying down.

3.3. Human shape model

For modelling individuals standing, a 2-D geometrical model is used. The head is figured by an ellipse, and the torso and the legs by two vertical rectangles (see figure 5). Such a model is simplistic, but its main advantage is that it avoids complex projections and evaluations in the image. For instance, one can take advantage of 2-D rectangles using integral images [16]. Let us recall that the aim here is to obtain a good approximation of the body occupancy in the image, in order to discriminate from other components in the image, such as humans lying on the ground, rather than accurately segment all the parts of the body and recover the exact attitude of individuals, which is out of the scope of this paper.

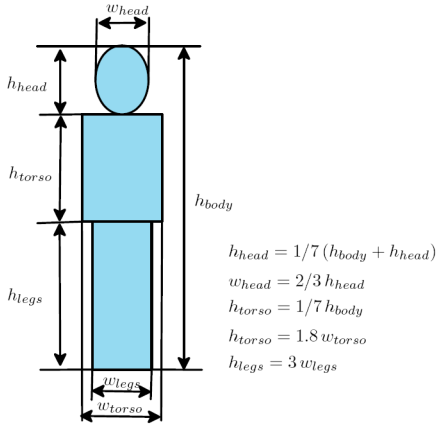

Figure 5. 2-D human shape model composed of an ellipse for the head, and two rectangles representing the torso and the legs.

3.4. Head detection

Following [17], head hypotheses are generated by two methods. The first one is a peak detection in the foreground image. A peak corresponds to a vertical local maxima in the extracted blobs. We use the region of interest definition as a very simple manual calibration procedure to have an estimate of a person size, given its horizontal position x in the image. The calibration also enables to determine the size of the window where local maxima are searched.

The second method is based on the detection of head shapes in the IR image using an elliptical template, as it was described in [2]. The idea is that in thermal images, highest gradients are often observed around exposed body parts, and especially for regions such as faces. For every position $\mathbf{x}$ of the template Γ center, a matching score is computed:

$$S_{\Gamma(\mathbf{x})} = \frac{1}{N_{\mathbf{x}_i}} \sum_{\mathbf{x}_i \in \Gamma(\mathbf{x})} \nabla I(\mathbf{x}_i) \cdot \mathbf{n}(\mathbf{x}_i) \qquad (1)$$

where $\mathbf{x}_i$ are distributed points on the ellipse, ∇I the gradient image, and $\mathbf{n}(\mathbf{x}_i)$ the normal to $\Gamma(\mathbf{x})$ at the point $\mathbf{x}_i$. Heads are searched only in the region of interest determined in 3.2 and the template scale is adapted to the horizontal position in the image.
Moreover, head hypotheses are validated by computing the intersection between the corresponding 2-D human shape and the foreground map F. To perform fast evaluations of this intersection, we use the integral image of F for the body model rectangular parts (torso and legs). All detections are finally merged with a sequential clustering algorithm.

3.5. Head-shoulder detection by a cascade of classifiers

Significant image gradients along the silhouette of individuals, and in particular the head-shoulders shape are relevant information to be used in the human detection. In addition to the previous method, we used the results of a head-shoulders detector based on local descriptors combined to build a boosted cascade[16]. Owing to the variability of appearance and pose that humans can have, we need a robust descriptor to represent relevant patterns. We worked with histograms of oriented gradients made of n orientations bins and 1 additional bin representing *the amount of information* inside the histogram support. They are computed densely, after luminance normalization, in position and scale on Regions of Interest (ROIs) of the image to capture finely the characteristics of the head-shoulders shapes we want to recognize. Our default parameters give 900 histograms. The integral image helps well for accumulating gradient values and votes and makes this computation efficient.

We noticed that this descriptor performs best with 9 bins of unsigned orientations (0°-180°with steps of 20°), thus we obtained vectors of 9000 components for each pattern. Unsigned orientations means that we do not make distinction between a dark-bright contrast and a bright-dark one (and this assumption makes sense with the variability of human appearances: hair, skin, clothes, etc.). To reduce aliasing, we smooth the histogram components by giving a fraction x of the vote to the corresponding bin and a fraction $1-x$ to the nearest bin where $x \in [x_{min}, 1]$. x_{min} depends on the angle threshold α_T above which we consider a vote belonging to one and only one bin ($x = 1$).

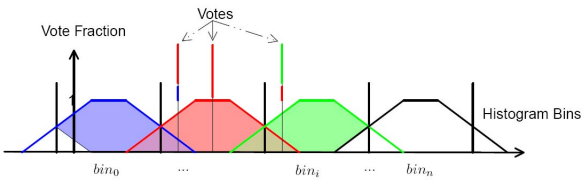

Figure 6. *Smooth* Histogram voting

The learning procedure is done by a cascaded boosting algorithm [6, 16] where the weak classifiers are simple decision stumps obtained from the histograms bins. This approach aims at decreasing as much as possible the number of candidate ROIs as we go further in the cascade, so that the first layer eliminates the majority of ROIs and the last layer of the cascade has only a few ROIs to evaluate. See figure 7 for a synthetic overview of this method. The evolution of the error and detection rates through the stages of the cascade enables the detector to reach good overall results. Indeed a 10 stages cascade can pretend to a detection rate of 0.9 if each stage has a detection rate of 0.99 (since $0.99^{10} \approx 0.904$). By the same way, a false positive rate of almost 10^{-4} is reached with 10 stages with a 40% false positive rate ($0.40^{10} \approx 1.0 \times 10^{-4}$). We tried several initialization parameters leading to cascades of 9 to 12 stages. The results of those detectors differ principaly on the final false rates and have almost similar detection rates.

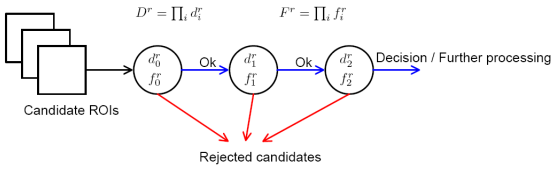

Figure 7. Cascade of classifiers, with D^r the detection rate and F^r the false positive rate.

For the learning procedure, we used $n_+ = 3000$ positive examples and $n_- = 10000$ *hard* negative examples. *Hard* examples are obtained from a sequence with a very simple detector (1 or few stages) trained on randomly choosen negative examples.

The boosting procedure selects relevant components – corresponding to the weak classifiers – to build a strong classifier which is then used to select $n-$ negative examples for the next stage of the cascade. The last layer trained is evaluated on a validation set and if it is not satisfying, a new layer is added to the cascade.

To improve this procedure and make it more robust, after the selection of weak classifiers for the current stage by the boosting algorithm, another loop adjusts the weights of these classifiers so that they are tuned more finely.

3.6. Segmentation refinement by MCMC sampling

Once the heads are accurately localized by our detector, we proceed with a segmentation which aim is to refine the position and the shape of the body models of people standing. This allows a better analysis of the remaining blobs after subtraction.
The segmentation problem can be formulated as a maximum a posteriori (MAP) estimation:

$$\Theta^* = \arg\max_{\Theta} p(\Theta|F) \tag{2}$$

where $\Theta = \{\theta_i\}$ is the set of the parameters of the human objects, and F is the foreground map given by the background subtraction. According to Bayes rule, the posterior probability $p(\Theta|F)$ can be decomposed into a likelihood term and a prior term:

$$p(\Theta|F) \propto p(F|\Theta)\, p(\Theta) \tag{3}$$

The parameters of each individual i are $\theta_i = \{\Delta x_i, h_i, f_i\}$, where Δx denotes the horizontal translation of the body with respect to its initial position, h the height, and f the fatness of the human model.

As the parameters of the various individuals are independent, we can assume the joint prior probability is the product of the prior probabilities of all the human objects:

$$p(\Theta) = \prod_{i=1}^{N} p(\theta_i) \tag{4}$$

where N is the number of detected people standing.
The prior probability of an individual i is:

$$p(\theta_i) = p(\Delta x_i)\, p(h_i)\, p(f_i) \tag{5}$$

$p(\Delta x_i)$ is a Gaussian distribution $\mathcal{N}(0, \sigma_{\Delta x})$ truncated in the range $[-0.4, 0.4]$, $p(h_i)$ is a uniform distribution in the range of $[h_{i0} - 0.3, h_{i0} + 0.3]$ (where h_{i0} is the initial size of the human model) and $p(f_i)$ is a uniform distribution in the range $[0.9, 2.2]$.
Since multiple humans may occlude each other, the joint likelihood cannot be decomposed into the product of likelihoods of individual human hypotheses. We use a joint likelihood based on the number of wrongly classified pixels, *e.g.* N_{01}, the number of pixels that belong to the foreground but are not within any body hypotheses, and N_{10}, the number of pixels that do not belong to the foreground but are within a body hypothesis:

$$p(F|\Theta) = \sigma(\lambda_{01}\frac{\Delta N_{01}}{N}) \cdot \sigma(\lambda_{10}\frac{\Delta N_{10}}{N}) \tag{6}$$

where is $\sigma(x) = 1/(1 + e^{-x})$ the sigmoid function, ΔN_{01} (resp. ΔN_{10}) is the difference between the current and the initial values of N_{01} (resp. N_{10}), and λ_{01} and λ_{10} are two coefficients depending on the mean size of a human in the image.
To maximize such a complex distribution, sampling methods are a simple way to explore the state space and find the optimal solution with good robustness to local maxima. The optimal parameters Θ^* are efficiently computed using a Markov Chain Monte-Carlo approach as in [17] for human segmentation or in [9] in the context of tracking multiple targets. The Metropolis-Hastings sampling algorithm is an efficient technique to draw samples from any probability distribution, by sequentially constructing a Markov chain that converges towards this distribution. We used a gaussian distribution as the proposal distribution. The main steps of the algorithm are:

- Initialize human shapes from head detections with parameters defined in 3.3 and scale determined by the horizontal position x in the image.
- For each sample:
  1. Randomly select an individual i,
  2. From the current state θ_i^t, predict a new state θ_i^{t+1} with the proposal distribution,
  3. Estimate the new posterior $p^{t+1}(\Theta|F)$,
  4. Calculate the acceptance ratio $r = \dfrac{p^{t+1}(\Theta|F)}{p^t(\Theta|F)}$
  5. If $r > 1$ the new state θ^{t+1} is accepted, otherwise it is accepted with the probability r

Once the samples from the posterior distribution are drawn, the state estimate is obtained by computing the weighted mean of the samples parameters.

4. Threat detection

Our threat detection system can output four possible answers for every new frame: *Empty* for a scene where no human was detected, *Normal* if individuals standing were detected and nobody is lying down, *Warning* or *Alarm*, depending on the level of confidence, in the case where people lying on the ground were detected. The decision is made by computing a threat detection probability associated to the *Lying Down* event $p_t(LD)$ and comparing it to two thresholds, a warning threshold τ_A and an alarm threshold τ_A such that $0 < \tau_W < \tau_A \leq 1$.

4.1. Lying down event detection

To determine whether there are people lying on the ground or not, the algorithm is based on:

- the height of detected heads: below $z_0 = 1\,m$, the person is classified as lying down (see 6 for little persons),
- the analysis of remaining blobs after the humans standing removal.

The segmented humans standing are removed from the foreground map. Morphological operations are then applied to clean the resulting binary map in order to eliminate thin regions. Human bodies usually present a high variability in texture do to their inhomogeneity in thermal images, in contrast to objects such as luggages for example. As a final step, the remaining blobs are processed and classified as laid persons or not, using the criterion of the distance from the ground and the texture information, characterized by the local variance computed in a neighbourhood of 3x3 pixels.

4.2. Threat probability

We note N_l the number of person lying down hypotheses at a given time t. The probability of a threat associated to the *Lying Down* event, $p_t(LD)$, can be expressed as a product of three probabilities. The first term $p_t^{N_l}(LD)$ is related to the number of detected people lying down hypotheses, the more hypotheses there are, the higher the probability of a threat is. The second probability $p_t^{X_l}(LD)$ depends upon the estimated position of the hypotheses with respect to the ground: the confidence level increases when the mean

distance to the ground becomes smaller. The third term $p_t^{f_l}(LD)$ expresses the frequency of the event detection in a temporal window. We keep a history of event detection in the temporal window of size h_w and count the number of occurences n_l of the LD event detection. The resulting probability can be written:

$$p_t(LD) = \underbrace{\frac{1}{1+e^{-\lambda_1 N_l}}}_{p_t^{N_l}(LD)} \cdot \underbrace{\frac{1}{1+e^{-\frac{\lambda_2}{N_l}\sum_{i=1}^{N_l}\frac{z_i}{z_0}}}}_{p_t^{X_l}(LD)} \cdot \underbrace{\frac{n_l}{h_w}}_{p_t^{f_l}(LD)} \quad (7)$$

λ_1 and λ_2 are two constant weighting parameters, they were experimentally set to $\lambda_1 = 1$ and $\lambda_2 = 2$. z_i stands for the distance from the hypotheses to the height threshold z_0.

5. Experimental results and discussion

The threat detector was extensively tested in the project. We present here results on two representative long sequences (3351 frames and 7342 frames respectively). The image dimensions are 384x272. In the sequences, a group of persons enters the emtpy area and stands on the platform. At a given time, some of them lie down, the other remain standing. The results obtained at the different stages of the algorithm are illustrated in the following figures. In figure 8, background subtraction results obtained with the SKDA method are presented. In all processed frames, no individual was missed in the foreground map. On the other hand, some false detections were observed, but they were filtered out in the next stages. Figure 9 shows results of head detection.

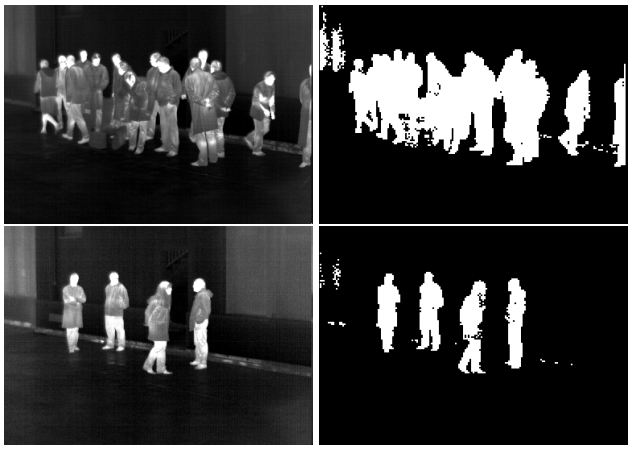

Figure 8. Background subtraction results.

The head detector was found to be very robust, given the complexity of the scenes in terms of local density of people, heavy occlusions and the variability of human postures. The elliptical shape detector and cascade classifier could detect heads while there was another individual standing behind, and even to some extent when a person is inclining. One can also notice the complementarity of the detectors for difficult cases. False detections below the height threshold were observed in only 14 frames over 10639 frames.

Figure 9. Results of head detection. Crosses indicate head found by local maxima and elliptical shape detector (in yellow above the height threshold, and in magenta below), boxes represent the head detection results obtained with the cascade classifier. The last picture is an example of false detection on legs which presents high gradients.

The number of head hypotheses above the height threshold is an estimation of the number of people standing in the scene. The accuracy of this estimation naturally depends of the density of people, and the level of overlapping. We plotted the number of head hypotheses generated for a sequence of 2555 frames. In that sequence, 15 persons are present and standing when the sequence starts. After approximately 1150 frames, 9 persons leave the area and during the last 600 frames, 6 individuals remain standing on the platform (see figure 10). The three main phases of the sequence are clearly visible on the plot. As it was expected, the estimation error is larger when the density of people increases, because of the occlusion effect, but the results are still consistent when compared to the ground truth.

The figure 11 shows results of the people standing segmentation with the Bayesian approach and the 2-D human shape model. The optimal shape parameters obtained after MCMC sampling enable to fit much better to the shape of individuals. In particular, a bending can be compensated by a translation of the body, and the height and the fatness are correctly estimated.

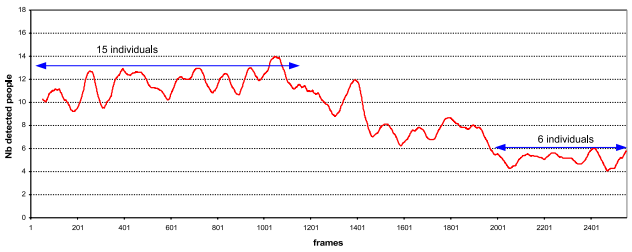

Figure 10. Estimation of the number of people standing. At the beginning of the sequence, there are 15 people standing in the area. 9 persons leave the area, 6 remain.

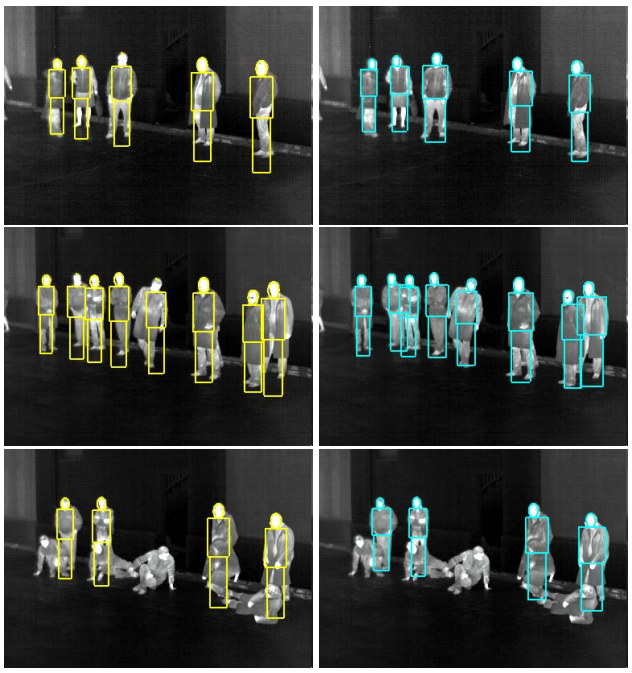

Figure 11. Results of human standing segmentation. Left column: initial position, right column: final segmentation after MCMC sampling.

After the people standing removal, remaining blobs detected near the ground are classified as people lying down or other object, as shown in figure 12. No detection cases are related to a high local density of people and extreme overlapping.

The table 1 gives the threat detection results: frames are classified as *Empty*, *Normal*, or *Warning/Alarm*. The algorithm gave satisfactory results, as the number of frames for each class in the estimation is consistent with the number of frames reported in the ground truth.

From the temporal point of view, if we plot the alarms over time (figure 13), we observe a good match between the algorithm output and the ground truth for the processed sequences. Note the very short delay of a few frames from the beginning of the threat annotated in the ground truth and

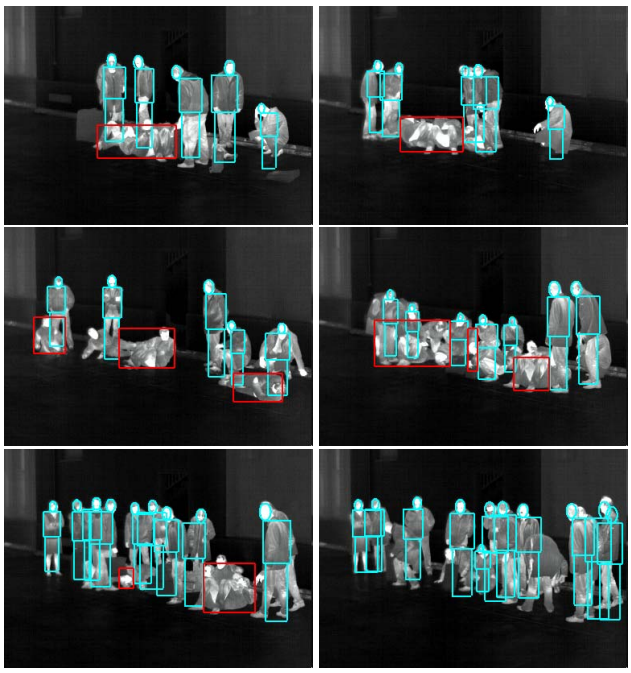

Figure 12. Detection of people lying down. The last image illustrates a case of no-detection.

Seq 1	*Empty*	*Normal*	*Warnings/Alarms*
Estimation	0	2214	1137
Ground truth	0	2161	1190
Seq 2	*Empty*	*Normal*	*Warnings/Alarms*
Estimation	0	4784	2557
Ground truth	0	4394	2948

Table 1. Threat detection results (in number of frames).

the first raised warning/alarm by our system, because of the size of the temporal window used to compute the frequency probability.

However, there are still some issues to improve. At the beginning of the sequence 2, false warnings/alarms appear very briefly. In addition, the alarm is not raised continuously during the critical period of time, because of intermittent cases of no detection of people lying down. The temporal smoothing of the threat detection output should be improved with a long-term filtering.

In terms of performance speed, with an unoptimized C++ code, the algorithm runs on a conventional PC Pentium IV 3Ghz, 1.5Gb RAM, at approximately 2-3 frames per second, which is an acceptable rate in the targeted application.

6. Conclusion and perspectives

In this paper, we demonstrated the capabilities of our system for analysing complex threat detection scenarios

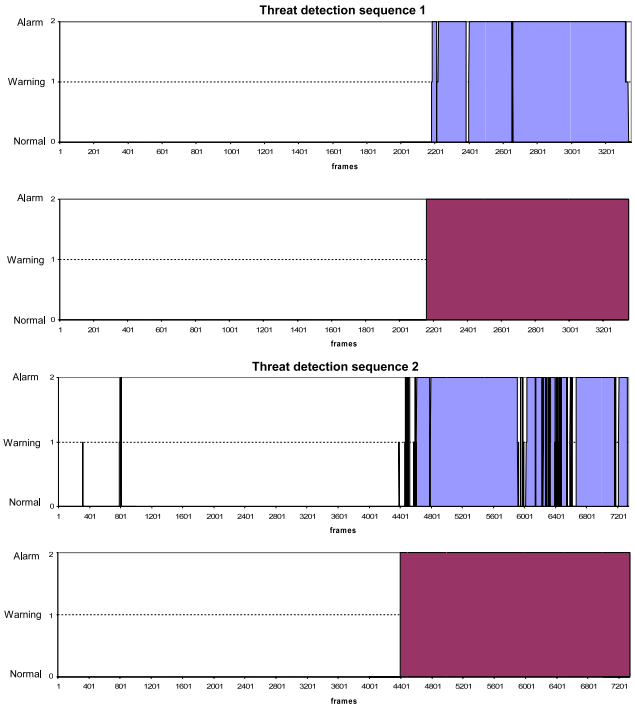

Figure 13. Threat detection results. First row: algorithm output for sequence 1, second row: ground truth for sequence 1, third row: algorithm output for sequence 2, last row: ground truth for sequence 2.

in thermal imaging such as the detection of people lying down in a crowded environment. The experimental results showed the robustness of the method, since the false alarm rate was low, and few of the expected alarms were missed. The performance of the detection could be improved by integrating a temporal smoothing, into both the segmentation process (visual tracking) and the threat detection output (long-term). Another improvement consists in enriching silhouette models to increase the segmentation accuracy with a limited computational cost. A finer model should enable to discriminate little persons from knelt ones. A multi-layer modelling of groups could also help us handle occlusions. Moreover, the results obtained in this study depend largely upon the position of the IR sensor and its field of view, which are related to current technology constraints. Ideally, a camera down-view with a wider FOV would minimize overlapping between individuals.

References

[1] E. L. Andrade, S. Blunsden, and R. B. Fisher. Hidden markov models for optical flow analysis in crowds. In *Proc. IEEE Int. Conf. on Pattern Recognition*, volume 1, pages 460–463, Los Alamitos, CA, USA, 2006.

[2] S. Birchfield. Elliptical head tracking using intensity gradients and color histograms. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pages 232–237, 1998.

[3] C. O. Conaire, E. Cooke, N. O'Connor, N. Murphy, and A. Smearson. Background modelling in infrared and visible spectrum video for people tracking. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition Workshop*, page 20, Washington, DC, USA, 2005.

[4] A. Davies, J. H. Yin, and S. Velastin. Crowd monitoring using image processing. *Electronics and Communications Engineering Journal*, 7(1):37–47, 1995.

[5] J. W. Davis and V. Sharma. Robust background-subtraction for person detection in thermal imagery. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition Workshop*, volume 8, page 128, Washington, DC, USA, 2004.

[6] J. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: a statistical view of boosting. Technical report, Dept. of Statistics, Stanford University, August 1998.

[7] B. Han, D. Comaniciu, and L. Davis. Sequential kernel density approximation through mode propagation: applications to background modeling. In *Proc. of the 2004 Asian Conference on Computer Vision*, 2004.

[8] B. Han and L. Davis. On-line density-based appearance modeling for object tracking. In *Proc. IEEE Int. Conf. on Computer Vision*, pages 1492–1499, Washington, DC, USA, 2005.

[9] Z. Khan, T. Balch, and F. Dellaert. Mcmc-based particle filtering for tracking a variable number of interacting targets. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 27(11):1805–1918, 2005.

[10] P. Kilambi, O. Masoud, and N. Papanikolopoulos. Crowd analysis at mass transit sites. In *ITSC '06, Intelligent Transportation Systems Conference*, pages 753 – 758, Toronto, Canada, September 2006.

[11] A. Marana, S. Velastin, L. Costa, and R. Lotufo. Automatic estimation of crowd occupancy using texture and nn classification. *Safety Science*, 28(3):165–175, 1998.

[12] D. L. Nanda H. Probabilistic template based pedestrian detection in infrared videos. In *IEEE Intelligent Vehicle Symposium*, pages 18–20, June 2002.

[13] A. Policastro and S. Gordon. The use of technology in preparing subway systems for chemical/biological terrorism. In *Commuter Rail/Rapid Transit Conference Proceedings*, 1999.

[14] P. Reisman, O. Mano, S. Avidan, and A. Shashua. Crowd detection in video sequences. In *IEEE Intelligent Vehicles Symposium*, pages 66–71, June 2004.

[15] G. T. and T. M.M. Pedestrian collision avoidance systems: A survey of computer vision based recent studies. In *ITSC '06, Intelligent Transportation Systems Conference*, pages 976–981, Toronto, Canada, September 2006.

[16] P. Viola and M. Jones. Robust real-time object detection. In *International Workshop on Statistical and Computational Theories of Vision Modeling, Learning, Computing and Sampling*, July 2001.

[17] T. Zhao and R. Nevatia. Bayesian human segmentation in crowded situations. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pages 459–466, June 2003.