Towards a sensor for detecting human presence and activity

Yannick Benezeth, Hélène Laurent, Bruno Emile, Christophe Rosenberger

Towards a sensor for detecting human presence and activity

Y. Benezeth[a], H. Laurent[b,*], B. Emile[c], C. Rosenberger[d]

[a] *Orange Labs, 4 rue du Clos Courtel, 35510 Cesson-Sévigné, France*
[b] *ENSI de Bourges, Institut PRISME, 88 bd Lahitolle, 18020 Bourges Cedex, France*
[c] *Institut PRISME, Université d'Orléans, 2 av F. Mitterrand, 36000 Châteauroux, France*
[d] *Laboratoire GREYC, ENSICAEN - Université de Caen - CNRS, 6 bd du Maréchal Juin, 14000 Caen, France*

Abstract

In this paper, we propose a vision-based system for human detection and tracking in indoor environment allowing to collect higher level information on people activity. The developed presence sensor based on video analysis, using a static camera is first of all presented. Composed of three main steps, the first one consists in change detection using a background model updated at different levels to manage the most common variations of the environment. A moving objects tracking based on interest points tracking is then performed. The classification step finally relies on the use of statistical tools and multiple classifiers for the whole body and for the upper-body. The validation protocol, defined by the industrial partners involved in the $CAPTHOM$ project focusing among other things on "Energy Management in Building", is then detailed. Three applications integrated into the $CAPTHOM$ draft finally illustrate how the developed system can also help in collecting useful information for the building management system: occupancy detection and people counting as well as activity characterization and 3D location extend to a wide variety of buildings technology research areas such as human-centered environmental control including heating adjustment and demand-controlled ventilation, but also security and energy efficient buildings.

Keywords: human detection, presence sensor, occupancy number detection, activity characterization, people localization

1. Introduction

The building sector is now one of those that consumes most energy. For example, in France, the building sector is responsible for 21% of the CO2 emission and for 43% of the total energy use. To economize energy, there are several solutions: first using renewable energies, second developing passive solutions such as insulation and third proposing solutions based on an active management of power consumption. This last approach requires to use reliable knowledge on buildings occupation. With this aim, we propose in this article a new sensor to detect human presence and to collect higher level information on people activity such as occupancy number detection and activity characterization.

Nowadays, the sensors available on the market are usually detectors whose technology is based on passive infrared. All living beings emitting heat, these sensors detect the electromagnetic radiations emit-

*Corresponding author.
Email addresses:
yannick.benezeth@orange-ftgroup.com (Y. Benezeth),
helene.laurent@ensi-bourges.fr (H. Laurent),
Bruno.Emile@univ-orleans.fr (B. Emile),
christophe.rosenberger@greyc.ensicaen.fr
(C. Rosenberger)

ted by humans of wavelengths between 6 and 14 μm. When a person moves in the detector field of view, the infrared radiation is focused by the Fresnel lens on the pyroelectric sensor. The corresponding moving hot spot causes the electronics connected to the chip to activate the detection. This technology is now well known and commonly used for lighting management, automatic door openers etc. However, it has several major flaws:

- motionless people can not be detected,
- the sensor is sensitive to shifts in air flow or sunshine radiations,
- the sensor is not able to distinguish between pets and humans.

The technological limits of these sensors, which are more motion than presence detectors, hinder the development of innovative solutions for energy consumption management. Conventional systems relying on a single occupancy sensor often suffer from a lack of data analysis of the measured sensor signals and cannot moreover differentiate between one or more occupants in the monitored space. In order to overcome these limits, several works have been conducted. They can mainly be gathered into two groups. The first one recommends the use of multiple low-cost, non-intrusive, environmental occupancy sensors, privileging the use of an independent distributed detectors network combined with a probabilistic data analysis. The second one applies for more expensive sensors such as video cameras.

The approaches belonging to the first group, that fuse together information from multiple sources, result in virtual sensors which are intended to be more powerful than single physical sensors [1, 2]. One proposal consists in combining, at occupied areas in the work space, three traditional inexpensive PIR occupancy sensors complemented with a sensor that determines when the telephone is off-hook [1]. If the use of probabilistic models offers improved capability of detecting occupant presence, the fundamental dependance on motion still remains. In order to address the problem of discerning the actual number of people in a room, complex sensor networks have been proposed. A second alternative, composed of a wireless ambient-sensing system complemented with a wired carbon dioxide sensing system and a wired indoor air quality sensing system, is considered to determine which parameters have the greatest correlation with the occupancy level [2]. The conducted tests show that, for the considered open office plan, CO_2 and acoustic parameters have the largest correlation with the number of occupants, complications arising however with acoustics because of the affect of sound by activities in nearby bays. Even if the proposed system achieves reasonable tracking of an actual occupancy profile, the achieved accuracy does not exceed 75% and, for certain days, it remains relatively low (e.g. 60%). Moreover, further exploration of sufficient training set sizes is needed.

Video cameras are also common especially when access to people activity or recognition are pursued [3, 4]. The main drawbacks of this solution are the need of large amounts of data storage and above all to interfere with privacy concerns. That is why several works propose to work with low-resolution cameras [5] or even develop "reduced" sensor from camera with a very different appearance from conventional video camera, allowing to obtain enough information to detect person's position and movement status but reducing the psychological resistance of having a picture taken [6].

Other approaches focus on occupancy location detection using ubiquitous and pervasive computing environments, often requiring non-ambient sensors such as wearable devices for tracking inhabitant activity [7]. If the major drawback of users psychological resistance to image capture can be overcome, the use of video camera remains the single sensor allowing to extract at the same time a wide range of information from low-level to high level interpretation. We can finally notice that most of the approaches privileging the use of multiple low-cost detectors exploit a camera network to establish the true occupancy information. The true occupancy numbers are then manually counted from the camera pictures [1, 2]. In order to at least facilitate the validation of new designed occupancy sensors, whatever can be the chosen technology, the design of an automated occupancy detector exploiting movies

is needed.

The works presented in this article take place within the *CAPTHOM* project which fits with the theme "Energy Management in Building" and aims at developing a reliable and efficient sensor to detect the presence of humans in indoor environments, using a static camera. The foreseen concerned sectors could be residential or tertiary areas. The main objectives of such a system are the power consumption management and the increase of comfort for residents, including for example heating adjustment considering their activity. Adaptability is also pursued. The *CAPTHOM* component must be easily integrated into other application areas related to security or supervision and assistance for elderly or disabled people, at home or in institutions.

The sensor must be able to detect the presence of a person in its environment without being disturbed by the presence of animals, other moving objects or by the morphology or activity of people. The sensor must be robust to light changes, heat sources and be able to detect people up to 15 meters. The complexity of the proposed algorithms, the used memory must be consistent with material constraints so that the treatment could be carried out in an embedded architecture. The developed system must therefore be an optimal compromise between false detection rate and algorithmic complexity. The proposed solution must be cheap, have a short time response and respect the European directive *2005/32/EC* of 6 July 2005 establishing a framework for the setting of eco-design requirements for energy-using products. Energy consumed by the sensor must be very low. Finally, the use of cameras imposes the respect of privacy and also implies the acceptance problem of this sensor by users. These issues were considered in the *CAPTHOM* draft and it was decided that the camera and its processing unit will only return useful information for the building management system. For the application, none image will be transmitted by the sensor to external devices.

We propose in this article, algorithmic solutions to detect people from image sequences. Several challenges must be settled :

- first, an image is a 2D representation of a 3D scene. The same object, observed from different views, may look very different,
- secondly, the image acquisition conditions may change from one environment to another. They can also vary all time long,
- thirdly, the backgrounds can be very complex. The possibilities of false detections are numerous and the contrast between people and background may possibly be very low,
- fourthly, many occlusions may appear between the person and the environment or among several individuals,
- finally, the main difficulty encountered for people detection is the very high intra-class disparity. Through their clothes, size, weight, outline, haircut etc., two individuals may appear very different. Moreover, the human body being highly articulated, the number of possible postures is great and the characteristics will vary in time.

In order to develop a system able to manage many of these situations, we tried to take into account all these difficulties. We developed a system using video analysis to interpret the content of a scene without doing strong assumptions about the nature of objects that could be present. Objects nature is determined by statistical tools derived from object detection. The next section describes the proposed algorithm using tools classically dedicated to object detection in still images in a video analysis framework. We then present various extensions of the project including its potential ability to recognize people activity. Finally, some conclusions and perspectives are given.

2. Presence sensor

In this section, we describe the human detection system proposed for the *CAPTHOM* project. It is based on video analysis obtained from a camera. This method has three main steps: change detection, moving objects tracking and classification. Figure 1 sums up the defined process for human detection.

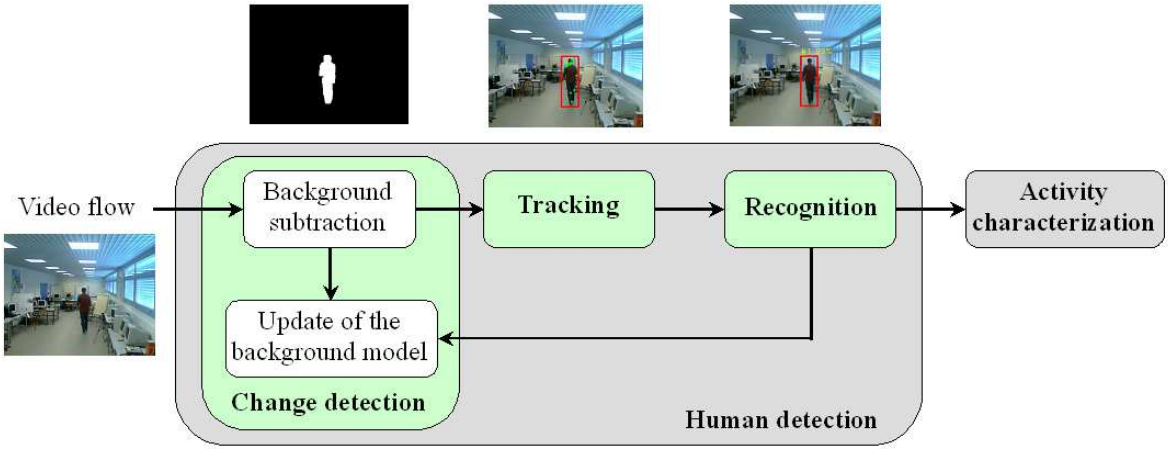


Figure 1: Process for human detection and activity characterization in video sequences. Illustration of the different intermediate results obtained all along the human detection chain: original image, change detection, tracking and recognition.

2.1. Change detection

Working with static camera, the search space can be reduced by detecting regions of interest in the image where there is a high probability to find a person. This first step is achieved through a background subtraction: from an environment model and an observation, we try to only detect what has changed.

According to the results obtained in the comparative study of background subtraction algorithms [8], we decided to model each pixel of the background by a Gaussian probability density function [9]. This quite simple model is a good compromise between quality of detection, computation time and memory requirement.

Since the scene is never completely static, the model must be designed to adapt to different environment changes, such as:

1. slow variation in lighting conditions, caused for example by natural change of daylights,
2. sudden and significant variation due for example to artificial extra lighting adding,
3. addition or removal of static objects.

The background model is consequently updated at three different levels: the pixel level updating each one with a temporal filter allowing to consider long time variations of the background, the image level to deal with global and sudden variations needing to reset the whole model and the object level to deal with the entrance or the removal of static objects. Performing these different updates at various levels, allows to manage the most common variations of the environment.

2.2. Tracking

After detecting the foreground pixels, the first step of tracking consists in gathering them into connected components, also called blobs afterwards, one blob ideally corresponding to one object (human or not) of the scene. Figure 2 presents an example of result which can be obtained after connected components gathering where one specific color has been attributed to each blob.

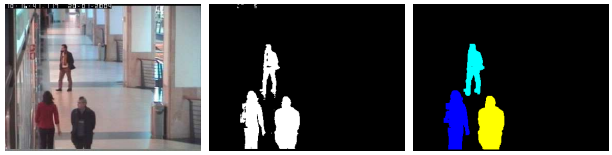


Figure 2: From the left to the right: original image, result obtained after background subtraction and finally after connected components gathering.

The objective of this step is to collect a history of movements for each blob. A blob potentially cor-

responding to one object, we would like to isolate each object of the scene and label it consistently over time. This history will be very useful to increase the overall system performances by smoothing in time misclassifications. In order to respect the constraints on computation time and to spare memory space, we decided to avoid complex model of tracking objects. The tracking process is then simply initialized with the connected components detected with the background subtraction. At every time t, we have to pair the list of detected blobs and the list of objects tracked in previous frames. To make the correspondence between these two lists, we use the matching matrix $\mathcal{H}_t$ defined as:

$$\mathcal{H}_t = \begin{pmatrix} \beta_{1,1} & \dots & \beta_{1,N} \\ \vdots & \ddots & \vdots \\ \beta_{M,1} & \dots & \beta_{M,N} \end{pmatrix} \quad (1)$$

where M is the number of tracked objects and N the number of blobs present at time t. If the tracked object i corresponds to the connected component j, $\beta_{i,j}$ is set to 1, otherwise $\beta_{i,j} = 0$. Each object is characterized by a set of interest points that are tracked frame by frame. The tracking of interest points is carried out with the pyramidal implementation of the Lucas and Kanade tracker [10, 11] with two additional constraints:

1. a tracked point must belong to the foreground; otherwise, the considered point is removed from the list of points and a new one is created,
2. when creating a new point, a distance constraint to other interest points is imposed in order to have an homogeneous distribution of points on any subject.

The correspondence between tracked objects and blobs is made if a minimum fixed percentage of interest points present on a blob is associated with an object.

The use of interest points also allows to manage some difficulties implied by our tracking method directly based on connected components. For example, when two distinct objects are very close, they form only one connected component and at the opposite, the same object can be represented by several blobs if there is a partial occlusion. The implemented tracking of interest points allows to deal with these common cases [12]. Figure 3 presents one example of results obtained after tracking. Partial occlusion is present when the two persons cross each other. We can observe from this example that the label of each person is consistent in time.

2.3. Recognition

Once regions of interest are detected and monitored, we have to detect their nature, namely if they correspond to a human or not. The classification method used in the proposed algorithm mainly relies on the one introduced by Viola and Jones [13], based on Haar-like filters and a cascade of boosted classifiers built with several weak classifiers trained with a boosting method [14].

Humans are detected in a sliding window framework. An area of interest is defined around the tracked object with a margin d on each side of its bounding box. This area of interest is analyzed by the classifier with various positions and scales. Figure 4 presents the bounding box of one detected object and the area of interest surrounding it. In a practical way, the classifier analyzes the area of interest with a constant shift in the horizontal and vertical directions. As the size of the person potentially present is not a priori known and the classifier has a fixed size, the area of interest is analyzed several times by modifying its scale. The size of the area of interest is divided by a fixed factor between two scales. By using a sliding window on several scales and positions, there are logically several overlapping detections which represent only one person. To fuse overlapping detections, we use, in order to have a fast computation, a simple average of results which intersect. The criterion of the Pascal competition [15] is used to detect the intersections. Figure 4 illustrates the multiple detections.

As in indoor environment, partial occlusions are frequent, it is clearly not sufficient to scan the area of interest only looking for forms similar to the human body in its whole. The upper part of the body (head and shoulders for example) is often the only visible part. It is possible to use a multiple parts representation of the human body in order to increase the

Figure 3: Illustration of tracking results with transient partial occultation. The first line corresponds to input images with interest points associated with each object (one color per object), the second line presents the tracking result with the label obtained for each tracked object.

robustness of the global system and manage partial occultation [16]. In practice, we used four classifiers:

1. the whole body (regardless of view point),
2. the upper-body (head and shoulders) from front or back views,
3. the upper-body (head and shoulders) from left view,
4. the upper-body (head and shoulders) from right view.

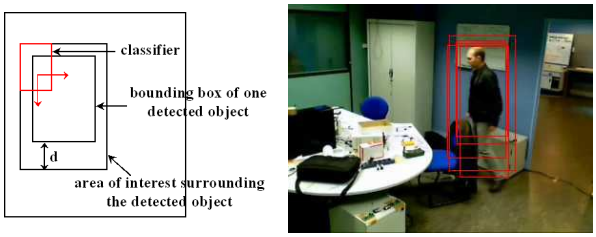


Figure 4: From the left to the right: search area analyzed by the classifier and resulting multiple detections.

It is important to use multiple classifiers for the upper body. Indeed, we empirically observed that a single classifier for the head and shoulders, under all points of view, was not sufficiently robust.

A confidence index $\mathcal{I}_{i,t}$ is then built for each tracked object i at time t depending on the confidence index obtained at previous frame and on confidence indexes corresponding to the detection result of each of the four classifiers. The final decision concerning the nature of the tracked object i is done using a simple thresholding.

3. Validation

In order to validate the proposed system, we first of all defined an experimental protocol including the definition of a test database, the implementation of reference algorithms to evaluate the relative performances of the proposed method and the choice of a comparative metric. These different sets and the obtained experimental results are presented in the next sections.

3.1. Test database

The consortium partners expressed their specific needs through a set of reference scenarios for which the sensor, namely the proposed algorithm, must respond appropriately. An extract of a scenario example is presented in figure 5. At each location, a set

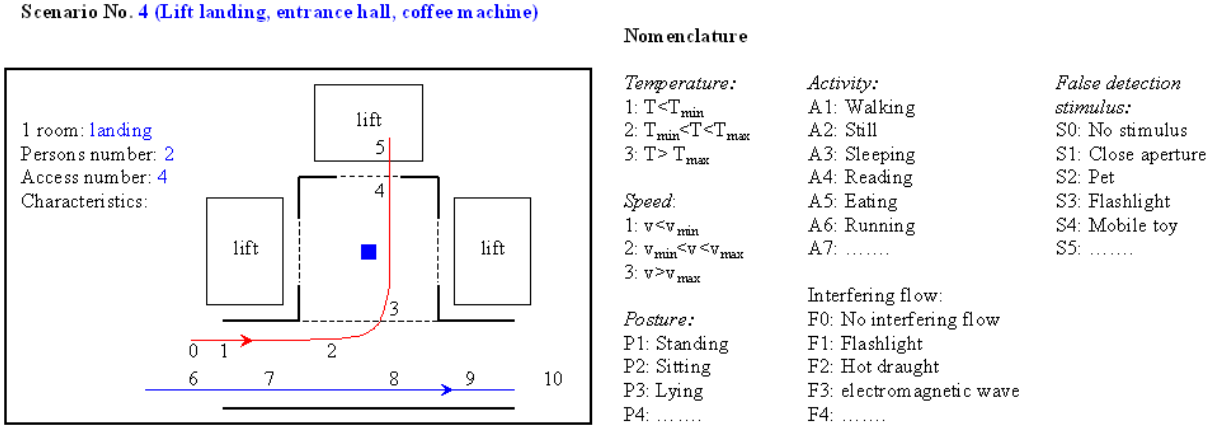


Figure 5: Extract of a scenario example defined by the industrial partners involved in the *CAPTHOM* project.

of charateristics (temperature, speed, posture, activity...) is associated with the formalism defined within the *CAPTHOM* project [17].

Three classes of scenarios constitute the evaluation database:

1. scenarios corresponding to a normal use of a room. In these scenarios, you can find one or more individuals that are static, moving, sitting or standing. These scenarios represent 14 videos realized in 9 different locations (offices, meeting rooms, corridors and dining rooms),
2. scenarios of unusual activities (slow or fast falls, abnormal agitation). 7 videos correspond to these situations,
3. scenarios involving all the stimuli of false detections identified by the consortium partners (variation of illumination, moving objects etc.). These situations represent 8 videos.

The total dataset is then composed of 29 videos taken in 10 different locations. The videos have a resolution of 320×240 and have an "average" quality since they were acquired with a simple webcam. Figure 6 presents examples of results corresponding to various scenarios.

3.2. Selected methods from the state of the art

In order to evaluate the interest of the proposed system to provide information on the presence or absence of human face to existing algorithms, we compared it with three reference systems:

- **IRP**: an existing person detector based on passive infrared technology,
- **Haar-Boost**: the detection system of Viola and Jones used with a sliding window scanning each image,
- **Haar-Boost + BS**: the previous method in which the search space of the classifier is reduced with a background subtraction.

3.3. Comparative metric

The results are presented using the confusion matrix (cf table 1), where:

- a is the number of correct detections (true positives),
- b is the number of missed detections (false negatives),
- c is the number of false detections (false positives),
- d is the number of true negatives.

Percentages are then computed.

Results extracted from a corridor scene with one or several individuals walking.

Results extracted from a meeting room scene with partial occlusions.

Results extracted from a laboratory scene with sudden illumination changes.

Figure 6: Examples of results corresponding to various scenarios.

	+	-
+	a	b
-	c	d

(1)

	+	-
+	$a/(a+b)$	$b/(a+b)$
-	$c/(c+d)$	$d/(c+d)$

(2)

Table 1: Example of confusion matrix (1) and presentation of results in percentages (2).

3.4. Experimental results

The performances obtained with the vision methods being highly dependent on the chosen scale factor, the algorithms have been tested with different values of this parameter. Only the best results for each method are presented here. Table 2 gathers the average results obtained on all scenarios.

We can first notice that the *IRP* detector presents quite low performances. Its main drawback is that it can only detect changes in temperature and therefore the movement of people. This characteristic induces many missed detections (false negatives) for many scenarios. Then, with *Haar-Boost*, even if the results are significantly better, the outcomes do not make a sufficient gap with commercial detectors. The number of false negatives and false positives remains too high. With *Haar-Boost + BS*, the number of false positives is significantly reduced. Indeed, with this method, the background subtraction can reduce the classifier search space and consequently the possible false detections. Finally, with the proposed method, we access to a history of people movements. We are able to average in time detections and therefore de-

	+	-
+	0.40	0.60
-	0.31	0.69

IRP

	+	-
+	0.77	0.23
-	0.58	0.42

Haar-Boost

	+	-
+	0.77	0.23
-	0.03	0.97

Haar-Boost + BS

	+	-
+	0.97	0.03
-	0.03	0.97

Proposed method

Table 2: Confusion matrix obtained on all scenarios.

tect a person even if, at a given time, the classifier can not recognize a person (laying person, poor contrast).

These results illustrate the benefits of jointly using background subtraction and tracking. The performance of the proposed method is good with a detection rate of 97% and a false detection rate of about 3%.

4. Applications

While it is important to have some information on the presence or absence of persons in an environment, it is not necessarily sufficient and computer vision can also help in collecting other useful information for the management system. We present in this section some examples of applications integrated into the *CAPTHOM* draft.

4.1. People counting

Improved building operation with respect to energy management and indoor environmental quality will be possible with an accurate detection of building occupancy resolved in space and time. Occupancy-based switching for power management of office equipment is one of the goal explicitly mentioned in the *CAPTHOM* project. In order to develop a demand-controlled ventilation or heating, it can be interesting to access, through the proposed sensor, to the current occupancy number of a supervised area. This information can be derived from the *CAPTHOM* sensor.

As an example, figure 7 presents the occupancy results obtained on two videos: the first one corresponds to an office scene ("Office" video), the second one to a corridor scene ("Corridor" video). The x-axis corresponds to time samples and the y-axis is the number of occupants in the space. The blue line is the actual occupancy profile and the red line is the estimated number of occupants, simply derived from the *CAPTHOM* sensor through a temporal smoothing. The profiles illustrate that the estimations track changes in occupancy fairly well. The obtained results describe, with a small delay, the major changes in occupancy. For an occupancy-based control scheme, this smoothed behavior is sufficient because abrupt fluctuations of short duration are rather

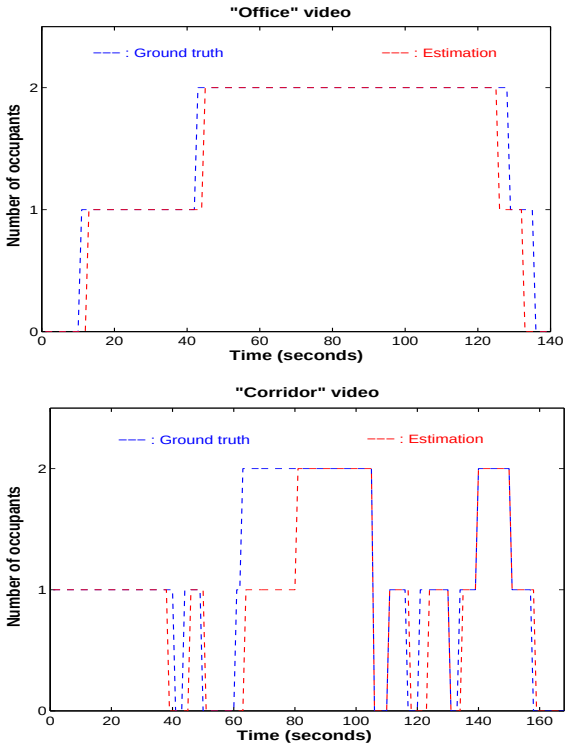


Figure 7: Occupancy estimation results on two videos respectively corresponding to an office and a corridor scene.

Results extracted from the "Office" video.

Results extracted from the "Corridor" video.

Figure 8: Examples of results extracted from the "Office" and "Corridor" videos.

insignificant. The total accuracy (number of correctly estimated points divided by the total number of points) is around 93% for the "Office" video and 83% for the "Corridor" video.

Both videos present one or two persons walking or sitting, with temporary partial occlusion. Interpreted images extracted from these two videos are presented in figure 8. The lower performances obtained for the second video are mainly due to the bad estimation done during the [60s-80s] period where only one person is detected while two are present. During this lapse of time, two persons enter the corridor in close succession, the first one occulting the second one quite entirely during about 20 seconds. With only one sensor, it is then impossible to distinguish one person from another. The second image presented in figure 8 for the "Corridor" video illustrates this situation. The use of two *CAPTHOM* sensors could allow to overcome this problem.

4.2. Activity characterization

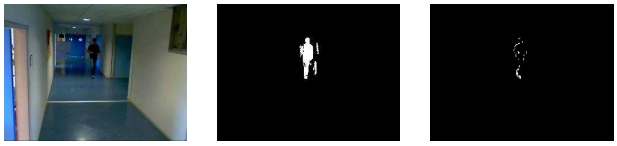
Easily derived from the *CAPTHOM* sensor, it is possible to access a simple quantification of people activity which could be useful, for example, to automatically regulate the heating. One can indeed imagine that the heat input that regulates the room temperature could differ depending on the number and activity of detected people. The foreseen activity characterization does not here include the semantic level of activity but simply a measure of restlessness.

The proposed test is based on the ratio between the number of pixels in motion and the number of pixels in the foreground. Since foreground pixels are available at each time t through the background subtraction step, we only need to calculate motion detection. Let $\mathcal{A}_{s,t}$ be the result of movement detection defined by:

$$\mathcal{A}_{s,t} = \begin{cases} 1 & \text{if } d_2(I_{s,t}, I_{s,t-\eta}) > \tau_{\mathcal{A}} \\ 0 & \text{otherwise,} \end{cases} \tag{2}$$

where $d_2(I_{s,t}, I_{s,t-\eta})$ represents the Euclidean distance between the pixel s at time t and the same pixel at time $t-\eta$ and $\tau_{\mathcal{A}}$ is a threshold. $\mathcal{A}_{s,t} = 1$ means that the pixel s is in a motion area. Figure 9 presents an example of foreground ($\mathcal{X}_{s,t} = 1$) and corresponding motion picture ($\mathcal{A}_{s,t} = 1$) obtained from a corridor scene. The measure of activity at time t, is performed by calculating the ratio between the number of pixels in motion and the number of pixels in the foreground:

$$\kappa_t = \frac{\sum_{s \in S} \mathcal{A}_{s,t}}{\sum_{s \in S} \mathcal{X}_{s,t}} \tag{3}$$

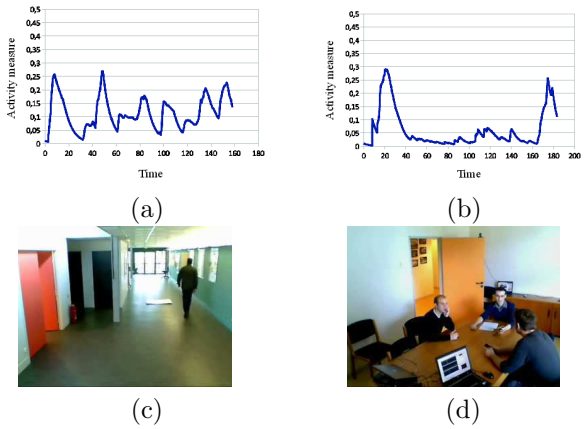


(a) (b) (c)

Figure 9: Examples of images used to measure the activity (a) input image (b) foreground (c) motion picture.

where S is the number of pixels in the image. A time filtering can be used to smooth over time the activity.

To validate this measure, we used this activity test on two groups of videos (extracted from the video dataset described in section 3.1). *Group 1* gathers 21 videos with low physical activity (stage office, meeting *etc*). *Group 2* gathers 6 videos with greater activity (corridor scenes or restless people). Figure 10 presents the evolution over time of the activity measure in two areas with distinct activities. The first case illustrates a corridor scene where each peak of the κ measure corresponds to the passage of a person. The second case corresponds to a meeting room where the two main peaks in the κ measure correspond to entrance and departure of people.



(a) (b) (c) (d)

Figure 10: Time evolution of the proposed activity measure in the case of a frequent passage zone (a) and (c) and in the case of a quiet meeting area (b) and (d).

Table 3 presents the results of activity measure into two classes "active" or "quiet". We can then observe that despite the simplicity of this measure, it is possible to obtain, with an interesting confidence level, a measure of the overall activity of humans. We get 91% of quiet response when the content of the scene is actually quiet and 100% of right answer for scenes with greater activity.

	quiet	active
quiet	0.91	0.09
active	0.00	1.00

Table 3: Confusion matrix over the classification of the scenes content between "quiet" and "active".

These results are of course dependent on the relative subjectivity of the activity concept. However, the proposed criterion of equation (3) allows to simply quantify the activity of people in a room. This measure of activity has two main advantages. First of all, it is very simple to calculate, especially since the foreground mask has already been calculated during the detection of persons. This criterion can be consequently easily implemented in an embedded hardware. Second, this measure does not depend on the camera point of view or on the distance between the humans and the camera. Despite of its simplicity, this information can be useful to automatically regulate heating.

4.3. People localization in a known environment

For home care applications, it may be interesting to know the position in space and posture of the detected individuals. Combining this information, we can detect abnormal situations: a person lying on his bed is a common situation, a person lying on the ground can certainly reflect a distress one.

First of all, two *CAPTHOM* sensors can be combined and mountained side by side. Each video stream of the stereovision system will then be independently processed based on the proposed algorithm before their results are merged together through a fusion step. Knowing the epipolar geometry of a stereovision system, each point observed in one image corresponds to an epipolar line in the other image. The so-called fundamental matrix can typically be used to determine such correspondance, the most commonly implemented method proposed to estimate it, being the 8-points algorithm [18]. The fundamental matrix

can also be obtained by combining the extrinsic and calibration matrices of each camera allowing to estimate the 3D position of each detected person. This information can reveal itself particularly interesting for the foreseen application.

It is then possible to locate humans in 3D and so to verify their activities. Figure 11 presents the estimated path of a person drawn into the floor map. In this figure, the individual enters from the right door, stays for a while near the shelf, and then leaves through the same door. We can note that this information is sufficient for home care applications where a more precise location would not be useful.

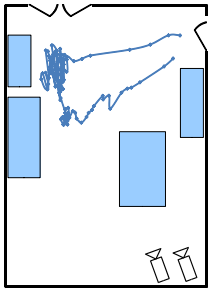


Figure 11: People localization and path estimation.

Possible improvements can be obtained combining different sensors. Most of the methods presented above were meant to work on daylight cameras (or visible). However, as their cost keep decreasing, far-infrared (FIR) cameras (often called IR or thermic cameras) gain more interest for human detection [19], [20], [21] as they provide numerous advantages (night vision, relatively uniform backgrounds, etc.). However, as opposed to daylight cameras, FIR cameras fail at detecting humans in hot summer days and often suffer from floor reflection. If one can link a point in one image to its corresponding epipolar line in the second image, this correspondence can also be used to confirm every human shapes detected in one camera with those detected in the other camera. Lets consider that M human shapes have been detected in the first image and N have been detected in the second image. We consider each top-left and bottom-right points of the i^{th} human shape (represented by a bounding box) in camera 1 and construct the respective epipolar lines in the camera 2 image using the fundamental matrix. As shown in figure 12, in our method, a detected shape $i \in [1; M]$ in camera 1 is kept if and only if there is a shape $j \in [1; N]$ such that the distance between the top-left and bottom-right points of the j^{th} shape and the corresponding projected epipolar lines is smaller than a predefined threshold (obtained empirically). In figure 12, two human shapes have been detected in camera 1 and only shape 1 has been kept. Of course, this algorithm is used both ways such that the shapes in camera 1 are confirmed with those in camera 2 and vice versa. Figure 13 presents on the first row, detections in a

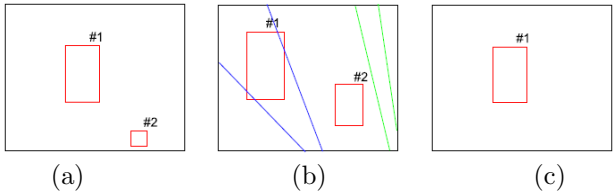


Figure 12: Example of detections fusion using the epipolar geometry (a) detections in camera 1 (b) detections in camera 2 and epipolar lines obtained related to camera 1 (c) detections in camera 1 kept after fusion.

FIR image and a daylight image with false positives. On the second row, detections have been fused with our method and false positives (the reflection on the floor on the left and the chair on the right) have been deleted. Precision/Recall curves obtained with this fusion process put into obviousness the improvement of human detection performances in both spectrums. Note that our fusion procedure is very fast since it only requires two projections per detection.

5. Conclusion

We have presented in this article a vision algorithm to detect human presence in an indoor environment. This algorithm combines background substraction, tracking and recognition. The evaluation of the proposed sensor, in many scenarios, gives a detection rate of 97%. The few remaining errors mainly come from three sources. A possible error is when a person is not detected. Through objects tracking, this case is relatively rare because the detections are smoothed

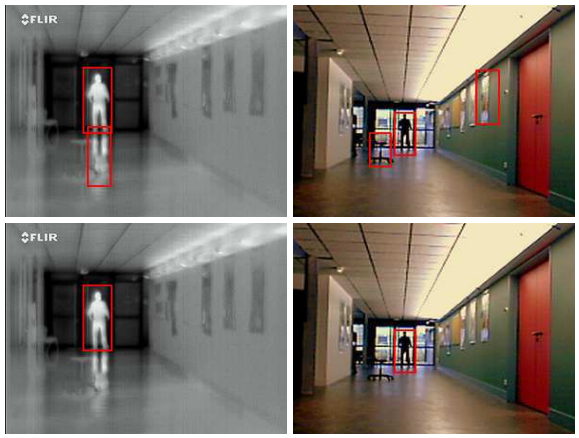


Figure 13: First row: detection examples without fusion step. Second row: detection examples with the stereovision system.

in time with the confidence index. However, it is possible that a person remains in a static configuration that the classifier does not recognize. These misdetections are mainly encountered when the contrast between the person and the background is not very distinctive or when the person takes a unusual posture. Then, when a person remains static for a long time, it is gradually included in the background model. If the setting speed of the background updating is properly fixed, this is not a problem. However, when the person moves again, there will be a "ghost" at the corresponding location. This ghost will be included in the background model through the object level update, but, the time lag between the departure of the person and the updated model of the background can lead to false detection. The third source of error directly comes from a possible failure of the background subtraction. Due for example to an illumination change of the scene, false detections can occur. Nevertheless, the different levels of background model updating restrict the appearance of such cases.

Some applications, including occupancy and activity characterization, have been presented. Initially intended for controlling office ambient conditions, many other developments can be envisaged. The vision can also be used to detect distress situations, such as the fall, without using a specific device but incorporating an additional module to the *CAPTHOM* sensor. An accurately known occupancy, both in time and space, can also allow to develop operational strategies for a better allocation of emergency personnel and resources in critical situations where every second counts. As part of the support, it can finally be useful to know the identity. Indeed, it may be advantageous to be able to distinguish between caregivers and patients. Similarly, to alert when there are people in places where they are not supposed to be, you need to know their identity. The use of cameras, with biometric technology, could be an interesting contactless solution for remote identification.

Acknowledgements

This work was made possible with the financial support of the Regional Council of Le Centre, the French Industry Ministry within the *CAPTHOM* project of the Competitiveness Cluster S2E2 (Sciences and Systems of Electrical Energy).

References

[1] R.H. Dodier, G.P. Henze, D.K. Tiller and X. Guo, "Building occupancy detection through sensor belief networks, " in *Energy and Buildings*, vol. 38, pp. 1033–1043, 2006.

[2] B. Dong, B. Andrews, K. P. Lam, M. Hynck, R. Zhang, Y.-S. Chiou and D. Benitez, "An information technology enabled sustainability test-bed (ITEST) for occupancy detection through an environmental sensing network, " in *Energy and Buildings*, 2010, doi:10.1016/j.enbuild.2010.01.016

[3] J. Han and B. Bhanu, "Fusion of color and infrared video for moving human detection, " in *Pattern Recognition*, vol. 40, pp. 1771–1784, 2007.

[4] P. Turaga, R. Chellappa, V. S. Subrahmanian and O. Udrea, "Machine Recognition of Human Activities: A Survey, " in *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, No. 11, pp. 1473–1488, 2008.

[5] I.J. Amin, A.J. Taylor, F. Junejo, A. Al-Habaibeh and R.M. Parkin, "Automated people-counting by using low-resolution infrared and visual cameras, " in *Measurement*, vol. 41, pp. 589–599, 2008.

[6] S. Nakashima, Y. Kitazono, L. Zhang and S. Serikawa, "Development of privacy-preserving sensor for person detection, " in *Procedia Social and Behavioral Sciences*, vol. 2, pp. 213–217, 2010.

[7] M. Chan, D. Estve, C. Escriba and E. Campo, "A review of smart homes - Present state and future challenges, " in *Computer Methods and Programs in Biomedicine*, vol. 91, pp. 55–81, 2008.

[8] Y. Benezeth, P.M. Jodoin, B. Emile, H. Laurent and C. Rosenberger, "Review and Evaluation of Commonly-Implemented Background Subtraction Algorithms, " in *Proc. International Conference on Pattern Recognition (ICPR)*, 2008.

[9] C. Wren, A. Azarbayejani, T. Darrel and A. Pentland, "Pfinder: Real-time tracking of human body, " in *IEEE Transaction on Pattern Analysis and Machine Intellingence (PAMI)*, vol. 19, pp. 780–785, 1997.

[10] B.D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision, " in *Proc. International joint conference on artificial intelligence*, pp. 674–679, 1981.

[11] J.Y. Bouguet, *Pyramidal implementation of the lucas kanade feature tracker: Description of the algorithm.* Technical report, Intel Corporation, Microprocessor Research Labs, 1999.

[12] Y. Benezeth, B. Emile, H. Laurent and C. Rosenberger, "Vision-based system for human detection and tracking in indoor environment," *Special Issue on People Detection and Tracking of the International Journal of Social Robotics (IJSR)*, vol. 2(1), pp. 41–52, 2010.

[13] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features, " in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2001, pp. 511–518.

[14] R.E. Schapire, "The boosting approach to machine learning: An overview, in *MSRI Workshop on Nonlinear Estimation and Classification*, 2002.

[15] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn and A. Zisserman, *The PASCAL Visual Object Classes Challenge 2008 (VOC2008) Results*, http://www.pascalnetwork.org/challenges/VOC/voc2008/workshop/index.html.

[16] K. Mikolajczyk, C. Schmid and A. Zisserman, "Human detection based on a probabilistic assembly of robust part detectors, " in *Proc. European Conference on Computer Vision (ECCV)*, LNCS, Vol. 3021, 2004, pp. 69–82.

[17] P. David, V. Idasiak and F. Kratz, "A Sensor Placement Approach for the Monitoring of Indoor Scenes, " in *Proc. European Conference on Smart Sensing and Context (EuroSSC)*, LNCS, Vol. 4793, 2007, pp. 110–125.

[18] R. Hartley and A. Zisserman, "Multiple view geometry in computer vision.", Cambridge University Press, 2000.

[19] M. Bertozzi, A. Broggi, A. Lasagni and M. Del Rose, "Infrared Stereo Vision-based Pedestrian Detection, " in *Proc. IEEE International Intelligent Vehicles Symposium (IVS)*, 2005, pp. 24–29.

[20] F. Xu, X. Liu and K. Fujimura, "Pedestrian Detection and Tracking With Night Vision," *IEEE Transactions on Intelligent Transportation Systemss (ITS)*, vol. 6, pp. 63–71, 2005.

[21] Y. Benezeth, B. Emile, H. Laurent and C. Rosenberger, "A Real Time Human Detection System Based on Far-Infrared Vision, " in *International Conference on Image and Signal Processing (ICISP)*, Lecture Notes in Computer Science, Volume 5099, 2008, pp. 273–280.