A Real Time Human Detection System Based on Far Infrared Vision*

Yannick Benezeth[1], Bruno Emile[1], Hélène Laurent[1],
and Christophe Rosenberger[2]

[1] Institut Prisme, ENSI de Bourges - Université d'Orléans - 88 boulevard Lahitolle,
18020 Bourges cedex - France
[2] Laboratoire GREYC, ENSICAEN - Université de Caen - CNRS, 6 boulevard
Maréchal Juin, 14000 Caen - France
yannick.benezeth@ensi-bourges.fr

Abstract. We present in this article a human detection and tracking algorithm using infrared vision in order to have reliable information on a room occupation. We intend to use this information to limit energetic consumption (light, heating). We perform first, a foreground segmentation with a Gaussian background model. A tracking step based on connected components intersections allows to collect information on 2D displacements of moving objects in the image plane. A classification based on a cascade of boosted classifiers is used for the recognition. Experimental results show the efficiency of the proposed algorithm.

1 Introduction

Vision based systems can nowadays be found in many applications, for monitoring goods in private areas or for managing security in public ones. Nevertheless, the relative robustness of vision algorithms, the camera miniaturization and the computation capacity of embedded systems permit other applications of vision based sensors. In order to keep at home low mobility people or to manage the energetic consumption, we need some reliable information on room occupation, the number and the activities of the house occupants. Vision based systems are probably the most efficient technology for this task.

The Capthom project, in which we are working, falls within this context. It consists in developing a human detection low cost sensor. This sensor will present advantages compared to existing sensors, that is to say, a strong immunity to intemperate detections and a great detection reliability. We want to have a reference platform which can give information on a room occupation. This platform will assess the performance of other affordable technologies (e.g. cameras in the visible spectrum). In spite of its prohibitive price, far infrared technology is the most convenient one to automatically watch a room. Acquisition is not influenced by the luminosity. In this framework, we have developed a far infrared algorithm which can detect and track a human in a room.

* This work was realized with the financial help of the Regional Council of Le Centre and the French Industry Ministry within the Capthom project of the Competitiveness Pole S^2E^2.

A. Elmoataz et al. (Eds.): ICISP 2008, LNCS 5099, pp. 76–84, 2008.

If the need of a reliable human detection system in videos is really important, it is still a challenging task. First, we have to deal with usual object detection difficulties (background variety etc.). Second, there are other specific constraints for human detection. The human body is articulated, its shape changes during the walk. Moreover, human characteristics change from one people to another (skin color, weight etc.). Clothes (color and shape) and occlusions also increase the difficulties.

Some human detection systems have already been proposed in the literature. First, some of them are outline based. They try to recognize the outline which is detected with a foreground segmentation. For instance, Kuno et al. [1] use the outline histogram for the classification. Dedeoglu [2] matches the detected outline with a outline from a database using a distance calculation. Mae et al. [3] compute a distance between the detected outline and a model. These methods are highly influenced by the foreground segmentation step. Moreover, they cannot deal with a moving camera and a crowded scene.

Other methods are based on machine learning. Papageorgiou et al. [4] have first proposed a detector based on Haar wavelets and SVM. Viola et Jones [5] have proposed a global framework for object detection based on the boosting algorithm and Haar wavelets. More recently, Dalal et Triggs [6] have proposed to combine the Histograms of Oriented Gradients with SVM. The good generalization capacity and performance of these techniques are well known. Nevertheless, the learning dataset quality is really important and difficult to set up. Moreover, for video surveillance applications, many available information (e.g. motion and past events) are not used in these methods.

In our project framework, we need a real-time and reliable algorithm. Machine learning algorithms are expensive in computation time but present good performances. So, from this observation, we propose in this article an extension of these machine learning algorithms using advantages given by the video. In our approach, the foreground segmentation is used in order to limit the search space of our classifier. As we don't use the contour of the extracted area for the classification, we are less dependent on the foreground segmentation than outline based methods. Moreover, the 2D tracking system improves the global performance because we have multiple images of the same person at different moments.

Each step of the proposed algorithm (cf. figure 1) will be detailed in this article. First, we apply a foreground segmentation method to localize moving blobs which have a higher temperature than the scene average temperature. Second, after a connected components clustering and a filtering, regions of interest are

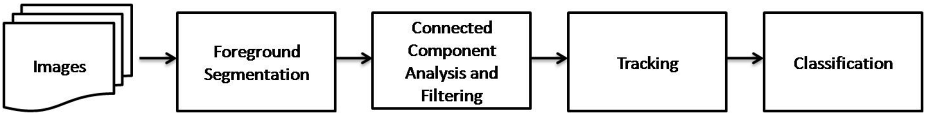

Fig. 1. Complete algorithm process

tracked to have chronological information of 2D movements. Then, we try to identify these connected components (that is to say if they are human beings). The performance of the proposed algorithm is shown in the last section with some experimental results.

2 Foreground Segmentation

We choose to first perform a foreground segmentation. The main objective is to simplify the image to proceed without any important information alteration. Only the detected regions of interest will be treated in the following steps. By definition, the background is the union of all static objects and the foreground is composed of all regions in which there is a high probability to find a human. There are two approaches for the foreground segmentation. Some algorithms are based on temporal differencing between two or three successive frames. Other algorithms perform a subtraction between the current frame and a background model. The most used background models are a temporal average, a single Gaussian distribution [7,11], a mixture of Gaussian [9] or a minimum and a maximum [8].

We have shown in [10] that a background subtraction with a single Gaussian distribution presents good performances in terms of detection and computation time. With this model, we take into account the acquisition system noise. As we are working in an indoor environment, we do not have to deal with a multimodal background model. We used this model in our algorithm, computing the average and the variance on each pixel. The foreground detection is realized by the following decision criterion :

$$\begin{cases} B_{1,t}(x,y) = 1 & \text{if } \; |I_t(x,y) - \mu_t(x,y)| > \tau_1.\sigma_t(x,y) \\ B_{1,t}(x,y) = 0 & \text{otherwise} \end{cases} \tag{1}$$

where (x, y) are the pixel coordinates and $I_t(x, y)$ its value in gray scale at time t ; $B_{1,t}$ is the binary image of the foreground detection ; μ_t and σ_t are respectively the average and the standard deviation and τ_1 is a threshold set empirically to 2.5.

If $B_{1,t}(x, y) = 0$, the Gaussian model is updated with :

$$\mu_t(x,y) = (1-\alpha).\mu_{t-1}(x,y) + \alpha.I_t(x,y) \tag{2}$$

$$\sigma_t^2(x,y) = (1-\alpha).\sigma_{t-1}^2(x,y) + \alpha.(I_t(x,y) - \mu_{t-1}(x,y))^2 \tag{3}$$

where α is a threshold determined empirically. Far infrared vision allows to see in night environment and gives also information about the scene temperature. With the hypothesis that the human temperature is noticeably higher than the average of his environment, we perform a binarization to detect hot areas in the image.

$$\begin{cases} B_{2,t}(x,y) = 1 & \text{if } \; I_t(x,y) > \tau_2 \\ B_{2,t}(x,y) = 0 & \text{otherwise} \end{cases} \tag{4}$$

where $B_{2,t}$ is the binary image representing hot areas, τ_2 is an arbitrary threshold. We finally apply a ”logic and” between the background subtraction and the hot area detection result :

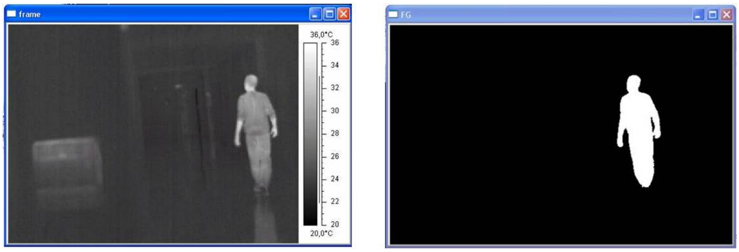

Fig. 2. Foreground segmentation example

$$B_t(x,y) = B_{1,t}(x,y) \cap B_{2,t}(x,y) \tag{5}$$

where B_t is the result of our foreground segmentation. On each frame, we gather detected pixels in connected components, these components are filtered deleting small ones. An example of foreground segmentation result is presented in figure 2.

3 Moving Objects Tracking

Once we have detected and filtered regions of interest in the image, we try to have information about blobs movements. In order to satisfy the real time constraints of our system, we have developed a relatively simple and fast algorithm based on the intersection of the connected components between frames at time t and at time $t-1$. We compute the matching matrix H_t at time t :

$$H_t = \begin{pmatrix} \beta_{1,1} & \dots & \beta_{1,N} \\ \vdots & \ddots & \vdots \\ \beta_{M,1} & \dots & \beta_{M,N} \end{pmatrix} \tag{6}$$

where M and N are, respectively, the components number of frames at time $t-1$ and at time t. $\beta_{i,j} = 1$ if the i^{th} component at time $t-1$ and the j^{th} component at time t intersect, otherwise $\beta_{i,j} = 0$. The analysis of the matrix H_t gives some information on the matching between the components at time $t-1$ and at time t. For example, if two components a and b at time $t-1$ and one component at time t intersect we merge components a and b in a new component.

We are able to deal with the merging or the splitting of blobs. However, we do not use any model for our tracked blob and we do not estimate the movement, we are not able to deal with occlusion problems. But, for our application, there are no consequences because, if an objet disappears, it is detected and considered as a new object when it appears again.

4 Human Recognition

Once we have chronological information of blobs movements, the following step is to know if the tracked blob is a human. To do that, we build a statistical model with a machine learning approach. There are many features and learning techniques in the literature. We have chosen the face detection system initially proposed by Viola et Jones [5]. This method is based on Haar wavelets and the boosting algorithm Adaboost.

Our learning dataset is composed of 956 positive images (cf. figure 3) and 3965 negative images (cf. figure 4). Images come from the OTCBVS dataset [12,13] (a benchmark database in the literature) and from images collected with an infrared camera where the ground truths is manually built. Because of the large number of negative images required, we use infrared images but also gray level images in the visible spectrum.

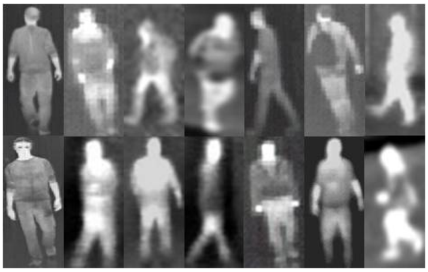

Fig. 3. Example of positive images

Fig. 4. Example of negative images

We use 14 features (Haar-like filters) described in figure 5. Each one is composed of two or three black and white rectangles. The feature values x_i are calculated with a weighted sum of pixels of each component.

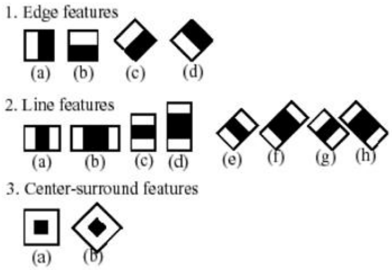

Fig. 5. Haar-like filters

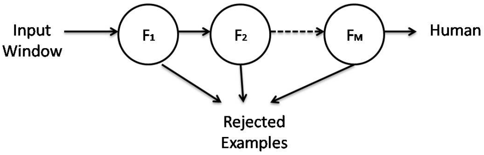

Fig. 6. Cascade of boosted classifiers

Each feature is then used as a weak classifier with :

$$f_i = \begin{cases} +1 & \text{if} \quad x_i \geq \tau_i \\ -1 & \text{if} \quad x_i < \tau_i \end{cases} \tag{7}$$

+1 means it corresponds to a human −1 not. Then, a more robust classifier is built with several weak classifiers using the boosting method [14].

$$F = sign(c_1 f_1 + c_2 f_2 + \ldots + c_n f_n) \tag{8}$$

We build a cascade of boosted classifiers (cf. figure 6). The neighborhood of the connected component tracked in the previous step is successively analyzed by each boosted classifier which can reject or accept the window.

5 Experimental Results

The speed of our detector is closely related with the number and the size of regions of interest. Approximatively, on a standard PC (Windows, 2GHz), for a $564*360$ frame size, our algorithm is able to proceed 30 frames per second when there is no blob in the image and 15 to 20 frames per second when there is at least one blob. A detection example is shown figure 7. The ellipse is the result of the blob detection, then we apply our detector in the neighborhood of this blob. A rectangle is drawn around the human if one is detected.

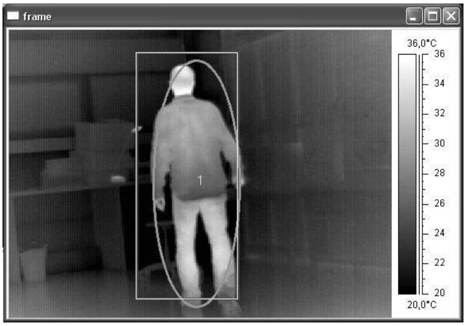

Fig. 7. A detection example, the rectangle is the human detection result in the blob neighborhood shown by the ellipse

We have also evaluated the performance of our classifier. We have built a dataset composed of 317 images. Each image contains at least one human, some infrared images are taken in indoor (room, office) and some in outdoor, some images contain pets.

The figure 8 shows a precision-recall curve obtained on this test dataset where :

$$\text{precision} = \frac{\text{right positives}}{\text{right positives} + \text{false positives}} \tag{9}$$

$$\text{recall} = \frac{\text{right positives}}{\text{right positives} + \text{false negatives}} \tag{10}$$

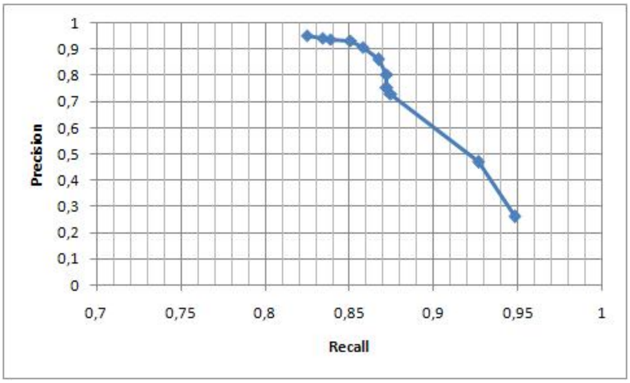

Fig. 8. Precision/Recall curve

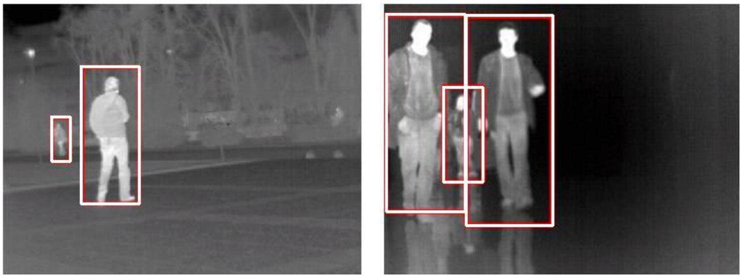

Fig. 9. Examples of right positive detections

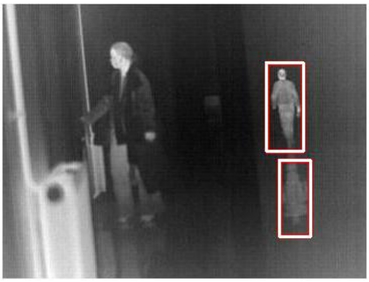

Fig. 10. Example of false positive and false negative detections

This curve has been obtained by varying the number of stages in the cascade. We obtain up to 95.34% for precision with 20 stages in the cascade and up to 94.78% for recall with 10 stages.

In figure 9, we show some examples of right positive detections. In figure 10, we show one example of false positive and false negative detections. In infrared vision, many surfaces are reflective, this is the cause of the majority of our false positive detections. Moreover, when someone wears a coat which is at ambient temperature, the distinction with the background is not easy. These two main difficulties are represented in figure 10.

Our classifier has been evaluated on static images but the global performance of our system will be improve with the tracking module.

6 Conclusion and Perspectives

Information on room occupation is really important for many systems, but human detection in video or in images is still a challenging task. In this article, we propose an extension of object detection systems using advantages given by the video. In our approach, the foreground segmentation is used in order to limit the search space of our classifier. Moreover, the 2D tracking system improves the global performance because we have multiple images of the same person at different moments.

Experimental results show the efficiency of our approach. Nevertheless, it remains several ways of improvement. First, the classifier performance is closely related to the database quality, and our database of infrared images can be improved. Second, we have to learn several classifiers for one human. As we work in indoor environment, occlusions are frequent, we could improve the robustness if we learn a part of the body which is more often visible (e.g. head and shoulder). A fusion with the visible spectrum can also decrease the number of false positive detections (because of the reflective surfaces in the infrared spectrum). Finally, we plan to develop our system in order to recover high-level information on human activities in a room.

References

1. Kuno, Y., Watanabe, T., Shimosakoda, Y., Nakagawa, S.: Automated Detection of Human for Visual Surveillance System. In: Proceedings of the International Conference on Pattern Recognition, pp. 865–869 (1996)
2. Dedeoglu, Y.: Moving object detection, tracking and classification for smart video surveillance, PhD thesis, bilkent university (2004)
3. Mae, Y., Sasao, N., Inoue, K., Arai, T.: Person detection by mobile-manipulator for monitoring. In: The Society of Instrument and Control Engineers Annual Conference (2003)
4. Papageorgiou, C., Oren, M., Poggio, T.: A general framework for object detection. In: 6th International Conference on Computer Vision, pp. 555–562 (1998)
5. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: Proceedings of the conference on Computer Vision and Pattern Recognition, pp. 511–518 (2001)
6. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 886–893 (2005)
7. Yoon, S., Kim, H.: Real-time multiple people detection using skin color, motion and appearance information. In: Proc. IEEE International Workshop on Robot and Human Interactive Communication, pp. 331–334 (2004)

8. Haritaoglu, I., Harwood, D., David, L.S.: W4: real-time surveillance of people and their activities. IEEE Transaction on Pattern Analysis and Machine Intelligence, 809–830 (2006)
9. Stauffer, C., Grimson, E.: Adaptive background mixture models for real-time tracking, CVPR, 246–252 (1999)
10. Benezeth, Y., Emile, B., Rosenberger, C.: Comparative Study on Foreground Detection Algorithms for Human Detection. In: Proceedings of the Fourth International Conference on Image and Graphics, pp. 661–666 (2007)
11. Han, J., Bhanu, B.: Detecting moving humans using color and infrared video. In: Proceedings of IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems, pp. 228–233 (2003)
12. Davis, J., Keck, M.: A two-stage approach to person detection in thermal imagery. In: Proc. Workshop on Applications of Computer Vision (2005)
13. Davis, J., Sharma, V.: Background-Subtraction using Contour-based Fusion of Thermal and Visible Imagery. Computer Vision and Image Understanding, 162–182 (2007)
14. Schapire, R.E.: The boosting approach to machine learning: An overview. In: MSRI Workshop on Nonlinear Estimation and Classification (2002)