A general framework for a robust human detection in images sequences *

Y. Benezeth[1], B. Emile[2], H. Laurent[1],C. Rosenberger[3]

[1]ENSI de Bourges
Institut PRISME
88 bd. Lahitolle
18020 Bourges, France

[2]Institut PRISME
IUT de l'Indre
2 av. F. Mitterrand
36000 Chteauroux, France

[3]GREYC, ENSICAEN
Université de Caen - CNRS
6 bd. Maréchal Juin
14000 Caen, France

Abstract

We present in this paper a human detection system for the analysis of video sequences. We perform first a foreground detection with a Gaussian background model. A tracking step based on connected components analysis combined with feature points tracking allows to collect information on 2D displacements of moving objects in the image plane and so to improve the performance of our classifier. A classification based on a cascade of boosted classifiers is used for the recognition. Moreover, we present the results of two comparative studies which concern the background subtraction and the classification steps. Algorithms from the state of the art are compared in order to validate our technical choices. We finally present some experimental results showing the efficiency of the proposed algorithm.

1 Introduction

The potential applications of a human detection system are numerous. We can for example quote systems used to monitor low mobility persons, home automation, video surveillance and event detections. For these applications, it is often necessary to detect humans before seeking highest level information. In our project, we attempt to develop a real-time system in order to limit the power consumption of buildings and to permit to monitor low mobility persons.

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 general object detection difficulties (background complexity, illumination conditions etc.). Second, there are other specific constraints for human detection. The human body is articulated, its shape changes during the walk. Then, human characteristics vary from one person to another (skin color, weight etc.). Clothes (color and shape) and occlusions also increase the difficulties.

There are two different approaches used to detect humans. On the one hand, there are methods which use an explicit model of the human shape (2D or 3D) (*e.g.* [1, 2]). On the other hand, and on the other hand there are methods based on machine learning technics. Based on a training database, these methods extract features (*e.g.* edges, gradients, shape, wavelet coefficients, etc.) and, following a clustering step (*e.g.* SVM, Adaboost, etc.), separate human from non-human shapes [3, 4, 5, 6].

Without any *a priori* knowledge, some classifiers use a sliding window framework, processed over the entire image, to detect humans. Whatever, some information can improve the global process. For example, for applications working with static cameras, it is possible to limit the search space of the classifier with background subtraction methods. Moreover, if a video flow is available, the temporal information can also be taken into account. Taking advantage of consecutive images presenting the same object at several moments, it will possible to follow each moving object independently and consequently to increase the classifier recognition rate.

In this paper, we present a human detection system in which the search space of the classifier is reduced calculating background subtraction with a single Gaussian model for the background. This choice is motivated through a comparative study of background subtraction methods presented in this article. Each object is independently tracked with a combination of connected components analysis and feature points tracking using SURF [7] features. Classification is done with Haar-like filters in a cascade of boosted classifiers [6]. We also present in this article a comparison of two well-known human detection classifiers [6, 4]. The global process is presented in figure 1

*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 Pole S^2E^2.

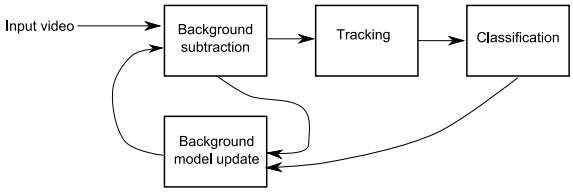


Figure 1. Overall process

2 Background substraction

As explained previously, background subtraction permits us to reduce the search space of the classifier, by localizing region of interest in the image. We first present a comparative study of background subtraction methods using real, synthetic and semi-synthetic videos. We then detail the chosen approach.

2.1 Comparative study

With the assumption that every moving object is made of a color (or a color distribution) different from the one observed in the background B, numerous background subtraction methods can be summarized by the following formula:

$$X_t(s) = \begin{cases} 1 & \text{if} \quad d(I_{s,t}, B_{s,t}) > \tau \\ 0 & \text{else,} \end{cases} \tag{1}$$

where X_t is the motion mask at time t, d is the distance between $I_{s,t}$ and $B_{s,t}$ respectively the frame and the background at time t and pixel s, τ is a threshold. Differences between all methods lie in the choice of the model used for B characterization and the distance d.

Basic motion detection method (Basic) The easiest way to model the background is to use a color image. This image could be obtained without moving object or estimated with a temporal median filter [8]. The distance between the background and the current frame could be done with:

$$d_2 = (I^R_{s,t} - B^R_{s,t})^2 + (I^G_{s,t} - B^G_{s,t})^2 + (I^B_{s,t} - B^B_{s,t})^2, \tag{2}$$

where R, G and B stand for the red, $green$ and $blue$ channels.

A Gaussian model (1-G) This method models each pixel of the background with a probability density function (PDF) determined by many learning frames. In this case, the subtraction of the background becomes a PDF thresholding problem. For instance, Wren *et al.* [9] model every background pixel with a Gaussian distribution $\eta(\boldsymbol{\mu}_{s,t}, \boldsymbol{\Sigma}_{s,t})$ where $\boldsymbol{\mu}_{s,t}$ and $\boldsymbol{\Sigma}_{s,t}$ stand for the average background color and covariance matrix at pixel s and time t. In this context, the distance metric can be the following Mahalanobis distance:

$$d_M = |\boldsymbol{I}_{s,t} - \boldsymbol{\mu}_{s,t}| \boldsymbol{\Sigma}^{-1}_{s,t} |\boldsymbol{I}_{s,t} - \boldsymbol{\mu}_{s,t}|^T. \tag{3}$$

Since the covariance matrix contains large values in noisy areas and low values in more stable areas, $\boldsymbol{\Sigma}$ makes the threshold locally dependent on the amount of noise.

A mixture of Gaussian model (GMM) To account for backgrounds made of animated textures (such as waves on the water or trees shaken by the wind), multimodal PDFs have been proposed. The most widely implemented approach is the Stauffer and Grimson's one [10] which models every pixel with a mixture of K Gaussians. For this method, the probability of occurrence of a color at a given pixel s is represented as:

$$P(\boldsymbol{I}_{s,t}) = \sum_{i=1}^{K} \omega_{i,s,t} . \eta(\boldsymbol{\mu}_{i,s,t}, \boldsymbol{\Sigma}_{i,s,t}) \tag{4}$$

where $\eta(\boldsymbol{\mu}_{i,s,t}, \boldsymbol{\Sigma}_{i,s,t})$ is the i^{th} Gaussian model and $\omega_{i,s,t}$ its weight.

Kernel Density Estimation (KDE) An unstructured approach can also be used to model a multimodal PDF. In this perspective, Elgammal *et al.* [11] proposed a Parzen-window estimate at each background pixel:

$$P(I_{s,t}) = \frac{1}{N} \sum_{i=t-N}^{t-1} K(I_{s,t} - I_{s,i}) \tag{5}$$

where K is a kernel (typically a Gaussian) and N is the number of previous frames used to estimate $P(.)$.

Minimum, Maximum and Maximum interframe difference (MinMax) The video surveillance system W4 [12] uses a model of the background composed of a minimum m_s, a maximum M_s and a maximum of consecutive frames difference D_s. A pixel s is in the background if:

$$|M_s - I_{s,t}| < \beta d_\mu \qquad \text{and} \qquad |m_s - I_{s,t}| < \beta d_\mu \tag{6}$$

where τ is a user-defined threshold and d_μ is the median of the largest interframe absolute difference over the entire image.

Presentation of the videos dataset To quantify the performance of the algorithms described above, these methods have been implemented on a wide range of real, semi-synthetic and synthetic videos. Our dataset is composed of 29 videos (15 reals, 10 semi-synthetics and 4 synthetics).

We created some synthetic and semi-synthetic videos, and others were downloaded from the databases PETS2001 [14] and IBM [15] and the competition VSSN 2006 [13]. Semi-synthetic videos are made with synthetic foreground (people, cars) moving on a real background. The whole dataset represents both indoor (20 videos) and outdoor scenes (9 videos). In addition, 6 videos contain animated background textures. Some examples of videos snapshots are shown in figure 2.

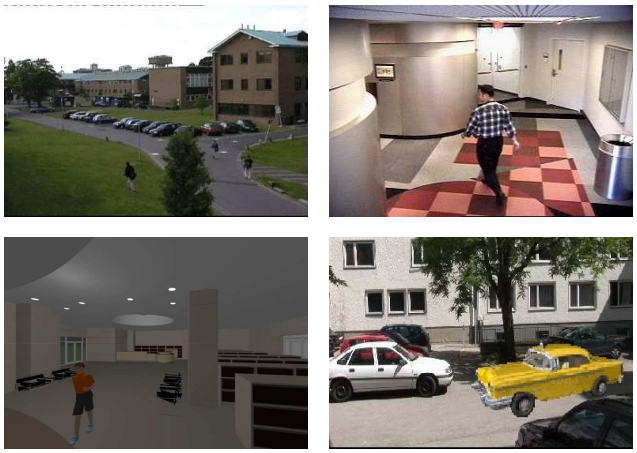


Figure 2. Snapshots of some videos in our dataset.

Results We perform the background subtraction algorithms described previously on various groups of videos illustrating specific situations: static (test 1), multimodal (test 2) and noisy (test 3) backgrounds. Results are presented using precision and recall values defined as follow:

$$Precision = \frac{RP}{RP + FP} \tag{7}$$

$$Recall = \frac{RP}{RP + FN} \tag{8}$$

where RP and FP represent the false positive and false negative detections. Results are presented in table 1.

For videos without noise with a perfectly static background (test 1), it is interesting to note that the results are globally homogenous. This is a rather interesting observation which means that simple methods such as *B*asic are as efficient as sophisticated ones when performing over simple videos. For multimodal videos (test 2), results are more heterogenous. Thanks to their multimodal shape, the *KDE* and *GMM* methods produce the most accurate results. Those obtained with the *1-G* method are surprisingly good. This can be explained by the fact that the *1-G* threshold is locally

	Basic	1-G	KDE	MinMax	GMM
$test\ 1$	0.92	0.92	0.88	0.88	**0.93**
$test\ 2$	0.55	0.76	**0.84**	0.48	0.79
$test\ 3$	0.62	**0.77**	0.75	0.28	0.76

Table 1. Recall values for different video sequences. $test\ 1$**: evaluation on noise-free with perfectly static background videos, precision is fixed to** 0.75**.** $test\ 2$**: evaluation on multimodal videos, precision is fixed to** 0.5**.** $test\ 3$**: evaluation on noisy videos, precision is fixed to** 0.75**.**

weighted by a covariance matrix which compensates well some background instabilities. Finally, with noisy videos (test 3), methods *1-G*, *KDE* and *GMM* produce good results with homogeneous recall values while the *MinMax* method does not seem appropriate for noisy videos. This can be explained by the fact that the *MinMax* threshold (which is global) depends on the maximum inter-frame difference(which is high for noisy videos).

Considering that sophisticated methods just slightly outperform simple ones in many cases, the choice of a simple background subtraction method is pertinent and sufficient.

2.2 Discussion

Following the comparative study presented in the previous section, we choose to model the background with a simple Gaussian model. The distance between the current image and the background model is done by calculating the Mahalanobis distance Eq. 3.

The background model is updated in three ways. Firstly, we apply a pixel-based update in which the mean and covariance of each pixel is iteratively updated following this procedure:

$$\boldsymbol{\mu}_{s,t+1} = (1-\alpha).\boldsymbol{\mu}_{s,t} + \alpha.\boldsymbol{I}_{s,t} \tag{9}$$

$$\boldsymbol{\Sigma}_{s,t+1} = (1-\alpha).\boldsymbol{\Sigma}_{s,t} + \alpha.(\boldsymbol{I}_{s,t} - \boldsymbol{\mu}_{s,t})(\boldsymbol{I}_{s,t} - \boldsymbol{\mu}_{s,t})^T. \tag{10}$$

Note that $\boldsymbol{\Sigma}$ is assumed to be diagonal to reduce memory and processing costs. Secondly, we update the background at the object-level. It means that, if the detected object is labelled as being non-human by the classifier with a sufficient confidence rate (presented further), we assigne the corresponding pixels to zero (which corresponds to the background). This permits us to avoid another useless processing. Thirdly, if a global change is detected due to a strong variation of illumination, *i.e.* if more than 80% pixels are

labeled as foreground, the background model is initialized with $B_t = I_t$ and $\Sigma_t = \sigma_0^2.Id.$

3 Tracking

Once we have detected regions of interest in the image, the following step consists in collecting information about objects displacements in the image plane. At each time t, we have m components detected by the background subtraction and n objects tracked at time $t-1$. We have based our method on the analysis of connected components detected by the background subtraction, regarding with the position of points of interest described with Speed Up Robust Features (SURF features) [7].

Thanks to Hessian-based detectors, points of interest are first detected over the current image for pixels located in the motion mask extracted at time t. Afterwards, the detected points neighborhoods are described with SURF features. Authors show in [7] that SURF features are faster to compute than SIFT and present good experimental results. Each interest point is consequently described with a descriptors vector of length 64. Then, the matching between points of interest at time t and $t-1$ is simply done by calculating the euclidian distance between their descriptors vectors. As we know that at time $t-1$, there were n tracked objects and that each descriptors vector corresponds to a particular object, we can say that the connected component at time t containing the majority of points of interest belonging to the i^{th} object at time $t-1$ is the most likely to correspond to the i^{th} object at time t.

4 Human recognition

The principle of humans recognition or more generally of objects recognition can be seen as a combination of descriptors and a classification method. The idea is to encode spatial information of an image into a descriptors vector and to use a learning technique for classification. In this paragraph, we first compare 2 widely-used classifiers: *Haar-Boost* (Viola and Jones [6]) and *HOG-SVM* (Dalal *et al.* [4]).

4.1 Comparative study

To use the best algorithm for our application, we tested *HOG-SVM* and *Haar-Boost* algorithms. These two methods are widely used for human detection, a description is given in the following:

HOG-SVM Histogram of Oriented Gradient descriptors provide a dense indeed overlapping description of image regions. The local shape information are captured by encoding image gradients orientations in histograms. Dalal *et al.* [16] have proposed Histogram of Oriented Gradients in the case of human detection combined with a linear SVM.

Haar-Boost Another approach widely used in the case of human detection is another dense and local representation: haar-like filters. Haar wavelets corfficients are used at different orientations and scale as a local feature. Viola and Jones [6] use Haar wavelets for object detection and then have extended their system in the case of human detection. More details are given in section 4.2

Presentation of the images dataset For the traininng, we use 1208 positive images (2416 images with vertical symmetry) and 3415 negative ones. Performances are compared on a dataset composed of 213 images containing at least one human standing with varied poses and postures.

Results Results are presented in figure 3. If we compare these algorithms based on their respective Precision/Recall curves, *Haar-Boost* is slightly better on our test dataset. Moreover, Haar-like filters are really fast to compute with the integral images.

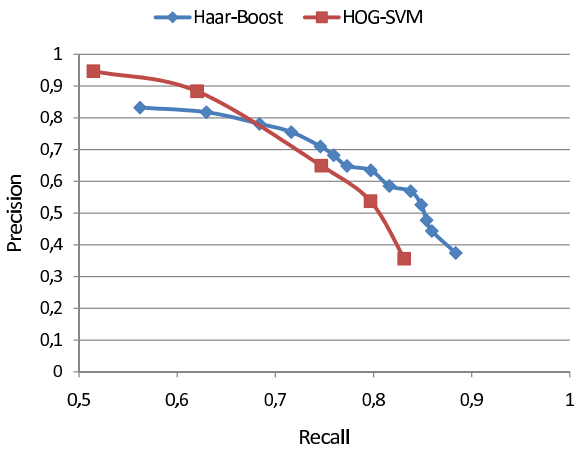


Figure 3. Comparison of *HOG-SVM* and *HAAR-BOOST*

4.2 Presentation of the classifier

In the *Haar-Boost* algorithm, 14 Haar-like filters are used and, as shown in Fig 4, those filters are made of two or three black and white rectangles. The feature values x_i are computed with a weighted sum of pixels of each component.

Each feature x_i is then fed to a simple one-threshold weak classifier f_i :

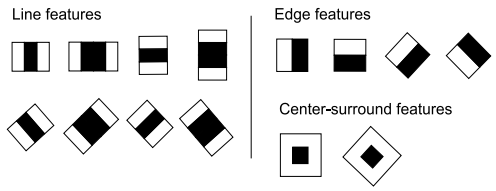


Figure 4. Haar-like filters used by the selected human recognition method.

$$f_i = \begin{cases} +1 \text{ if } x_i \geq \tau_i \\ -1 \text{ if } x_i < \tau_i \end{cases} \quad (11)$$

where $+1$ corresponds to a human shape and -1 to a non-human shape. The threshold τ_i corresponds to the optimal threshold that minimizes the misclassification error of the weak classifier f_i estimated during the training stage. Then, a more robust classifier is built with several weak classifiers trained with a boosting method [17]:

$$F_j = sign(c_1 f_1 + c_2 f_2 + \ldots + c_n f_n). \quad (12)$$

A cascade of boosted classifiers is built (cf. Fig. 5). F_j corresponds to the boosted classifier of the j^{th} stage of the cascade. Each stage can reject or accept the input window. Whenever an input window passes through every stages, the algorithm labels it as a human shape. Note that humans are detected in a sliding window framework [6].

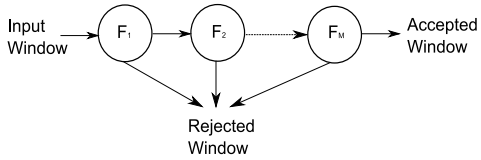


Figure 5. Cascade of boosted classifiers.

4.3 Temporal integration

As we independently track moving objects in the scene, we are able to build a confidence index about the nature of the moving objects. This index depends on the number of previous frames where the object is recognized as a human or not. It can be written as follows:

$$\Lambda_t = \begin{cases} \Lambda_{t-1} + \frac{100 - \Lambda_{t-1}}{\beta} & \text{if a human is detected} \\ \Lambda_{t-1} - \rho & \text{else} \end{cases} \quad (13)$$

where Λ_t is the confidence index at time t, β and ρ two tunable parameters.

5 Experimental results

We have previously presented quantitative results validating our technical choices for the background subtraction and the classification. To illustrate the global processus, we present some snapshots of different steps of our method in figure 6.

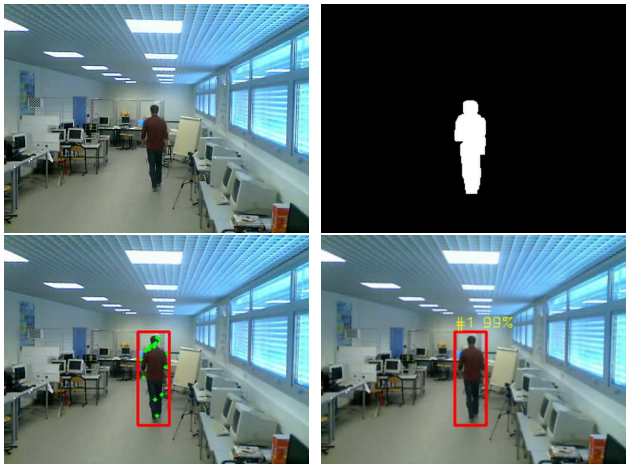
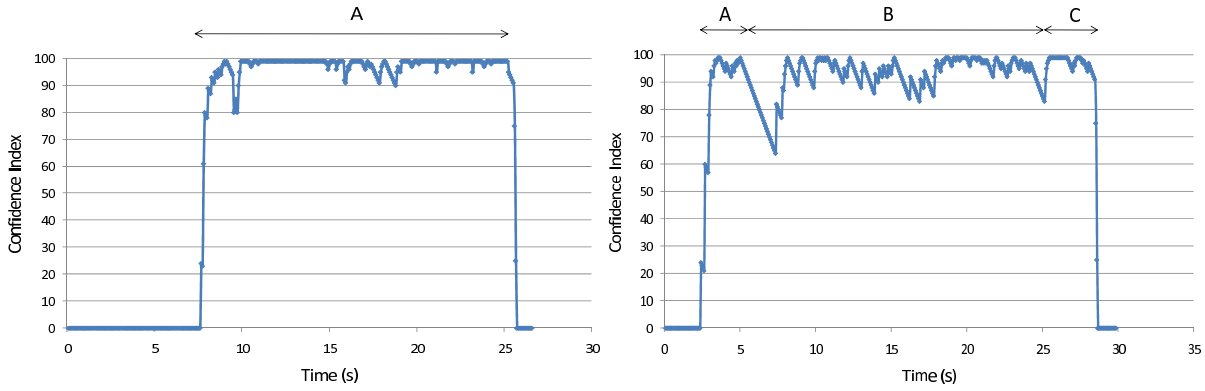


Figure 6. Illustration of the proposed method. From top-left to bottom-right: input image, result of the background subtraction, tracking and final result with the confidence index.

We also present the evolution of the confidence index for two different videos in figure 7. The figure 7-(a) represent a simple case in which one man is moving in the room. We can observe that the confidence index is very hight. The figure 7-(b) represent the confidence index for a case in which the person is partially occluded. The confidence index in this case is still very hight but present some irregularities.

6 Conclusion

In this paper, we present a real-time human detection system in which the search space of the classifier is reduced calculating background subtraction with a single gaussian model for the background. Each object is independently tracked with a combination of connected component analysis and points of interest matching using SURF features. Classification is done with Haar-like filters in a cascade of boosted classifiers. Two comparative studies of background subtraction methods and of two well-known human detection classifiers are presented motivating our technical choices. We first show that simple background subtraction methods are as efficient than complex methods in many cases and then we show that the Viola *et al.* present best



(a) (b)

Figure 7. Evolution of the confidence index. In (a), one man is in the room during A. In (b), one man is in the room, he is partially occluded during B.

results in our test dataset. We finally present some experimental results validating our approach.

In the future, we plan to carefully evaluate the human detection system with a *global evaluation* combining localization and detection performance. Finally, we plan to develop our system in order to recover high-level information on human activities in a room.

References

[1] Q. Zhao, J. Kang, H. Tao and W. Hua, "Part Based Human Tracking In A Multiple Cues Fusion Framework", *ICPR*, pp. 450–455, 2006.

[2] Liang Zhao, "Dressed Human Modeling, Detection, and Parts Localization", *PhD thesis, The Robotics Institute, Carnegie Mellon University, Pittsburgh*, 2001.

[3] C. Papageorgiou and T. Poggio, "A trainable system for object detection," *IJCV*, vol. 38, pp. 15–33, 2000.

[4] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," *CVPR*, vol. 2, pp. 886–893, 2005.

[5] N. Dalal, B. Triggs and C. Schmid, "Human detection using oriented histograms of flow and appearance", *ECCV*, vol. 2, pp. 428–441, 2006.

[6] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," *CVPR*, pp. 511–518, 2001.

[7] H. Bay, A. Ess, T. Tuytelaars, L. Van Gool, "SURF: Speeded Up Robust Features", *CVIU*, vol. 110, pp. 346–359, 2008.

[8] Q. Zhou, J. Aggarwal , "Tracking and classifying moving objects from video", *PETS Workshop*, 2001.

[9] C. Wren, A. Azarbayejani, T. Darrel, A. Pentland, "Pfinder: Real-time tracking of human body", *PAMI*, 1997.

[10] C. Stauffer and W.E.L. Grimson, "Adaptative background mixture models for real-time tracking'", *CVPR*, 1999.

[11] A. Elgammal, D. Harwood and L. Davis, "Non-parametric model for background subtraction", *ECCV*, 2000.

[12] I. Haritaoglu, D. Harwood and L.S. Davis, "W4: real time surveillance of people and their activities", *PAMI*, 2000.

[13] http://imagelab.ing.unimore.it/vssn06

[14] www.cvg.cs.rdg.ac.uk/pets2001

[15] L. Brown, A. Senior, Y. Tian, J. Vonnel, A. Hampapur, C. Shu, H. Merkl and M. Lu, "Performance evaluation of surveillance systems under varying conditions." *PETS Workshop*, 2005.

[16] N. Dalal, B. Triggs, C. Schmid, "Human detection using oriented histograms of flow and appearance", *ECCV*, 2006.

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