Detection de la Presence Humaine par Vision Infrarouge : Application a la Gestion de l'energie electrique dans l'habitat

Yannick Benezeth, Bruno Emile, Hélène Laurent, Christophe Rosenberger

Détection de la Présence Humaine par Vision Infrarouge : Application à la Gestion de l'Énergie Électrique dans l'habitat *

Yannick Benezeth[1], Bruno Emile[1], Hélène Laurent[1] et 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

Résumé. Nous présentons dans cet article un algorithme de détection et de suivi d'humains, basé sur la vision infrarouge. L'objectif est ici d'avoir des informations fiables sur l'occupation d'une pièce. Nous utilisons pour cela, une segmentation de l'avant-plan avec une modélisation de chaque pixel de l'arrière-plan par une distribution gaussienne, un suivi de cibles basé sur le recouvrement des composantes connectées et une classification basée sur une cascade de classifieurs boostés. Cet algorithme nous permet d'avoir un historique de l'occupation d'une pièce quelle que soit la luminosité. Les résultats expérimentaux montrent l'efficacité de l'algorithme proposé.

Mots Clés : Vision artificielle, détection d'humains, segmentation de l'avant-plan, suivi, classification.

Introduction

L'importance de la vision pour la surveillance des biens et la protection des personnes est aujourd'hui bien connue. Les performances des algorithmes actuels, la miniaturisation des caméras et les capacités de calcul des outils embarqués permettent d'entrevoir de nouvelles applications pour les technologies basées sur la vision. En particulier pour les systèmes d'aide au maintien à domicile et pour les systèmes de gestion de la consommation énergétique, il est nécessaire d'avoir des informations fiables sur la présence, le nombre et l'activité des personnes dans l'habitat. Alors que les performances des capteurs actuels (détecteurs de mouvement pyroélectrique...) ne sont en général pas suffisantes puisqu'ils ne peuvent détecter que des mouvements et non pas, à proprement parler, la présence, l'utilisation de systèmes basés sur la vision ouvre de nouvelles perspectives.
Le projet Capthom s'inscrit dans ce contexte. Il consiste à développer un capteur de présence humaine dans un habitat. Il devra présenter des nets avantages par rapport aux capteurs existants, c'est à dire une forte immunité aux détections intempestives et une grande fiabilité de détection (personnes immobiles). Nous souhaitons disposer d'une plate-forme de référence permettant d'établir l'historique de l'occupation d'une pièce. Nous nous sommes intéressés à la vision infrarouge. En effet, malgré un coût prohibitif, cette technologie est celle qui fournit le plus d'informations, est la moins sensible aux perturbations extérieures et surtout permet d'avoir une image de la scène la nuit. Cette plate-forme permettra par la suite de valider d'autres algorithmes de vision basés sur des technologies plus abordables (spectre visible ou proche infrarouge) et de quantifier les performances des capteurs développés pour le projet Capthom. Nous avons donc développé un algorithme de traitement d'images, basé sur la vision infrarouge, capable de détecter un humain dans une pièce et de fournir un historique de l'occupation de la pièce.
Même si la demande est forte, la détection d'un humain dans une image ou dans une vidéo est un problème qui reste aujourd'hui ouvert. Il y a tout d'abord des problèmes généraux, communs aux systèmes de reconnaissance de formes pour des applications réelles (variation des conditions d'acquisition). Il y a ensuite des contraintes spécifiques à la détection d'un humain dans une image. Tout d'abord, le corps est hautement articulé. La silhouette d'une même personne change au cours de la marche. Ensuite, les caractéristiques des

* Ce travail a été réalisé avec le soutien financier de la Région Centre et du Ministère de l'Industrie dans le cadre du projet Capthom du pôle de compétitivité S^2E^2

humains varient d'une personne à l'autre (couleur de la peau, coupe de cheveux, poids etc.). Les vêtements (la texture et la forme) et les possibilités d'occultation compliquent aussi grandement le problème.
Plusieurs approches ont été proposées dans la littérature pour détecter un humain dans une image ou une vidéo. Une première approche est basée sur la silhouette du corps humain, détectée par une segmentation de l'avant-plan. Après avoir extrait la silhouette du corps humain, Kuno et al. [1] utilisent l'histogramme de la silhouette pour la classification. Dedeoglu [2] compare la silhouette extraite avec une base de données en calculant une distance entre la silhouette détectée et chaque silhouette de la base. Mae et al. [3] calculent une distance entre la silhouette détectée et un modèle prédéfini. Ces méthodes sont très fortement dépendantes de la qualité de la segmentation de l'avant-plan. De plus, elles ne peuvent fonctionner avec des caméras en mouvement et des avant-plans denses.
On trouve également d'autres méthodes basées sur les différentes techniques d'apprentissage. Papageorgiou et al. [4] ont les premiers proposé un détecteur basé sur les ondelettes de Haar et les séparateurs à vaste marge. Viola et Jones [5] ont, quant à eux, proposé un système de détection basé sur l'algorithme du boosting et les ondelettes de Haar. Plus récemment, Dalal et Triggs [6] ont développé une méthode basée sur la combinaison des histogrammes de gradients orientés et des séparateurs à vaste marge. Les bonnes capacités de généralisation et les performances de tels systèmes sont aujourd'hui bien connu. Néanmoins, la base d'images utilisée pour l'apprentissage est primordiale et est assez lourde à mettre en place. De plus, pour les applications de vidéosurveillance, beaucoup d'informations (le mouvement, les événements passés) ne sont pas utilisés.
Dans cet article, nous proposons une extension des techniques basées sur l'apprentissage en utilisant les avantages offerts par la vidéo. Dans notre approche, la segmentation de l'avant-plan est utilisée pour limiter l'espace de recherche de notre classifieur. Comme nous n'utilisons pas la forme de la silhouette détectée, nous sommes moins dépendant de la segmentation de l'avant-plan que les méthodes où la silhouette est utilisée pour la classification. De plus, le suivi de cibles 2D augmente aussi les performances globales parce que nous avons plusieurs images d'une même personne à différents instants.
Chaque étape du processus (cf. figure 1) est détatillé dans cet article. Une segmentation de l'avant-plan est d'abord réalisée pour localiser les objets en mouvement qui ont une température supérieure à l'environnement dans l'image. Dans un second temps, nous regroupons les composantes connectées et nous les filtrons (les composantes trop petites sont supprimées). Les zones d'intérêts mises en relief dans les étapes précédentes sont suivies, trame par trame, afin d'obtenir un historique des déplacements 2D. Ensuite, sachant la position de l'objet d'intérêt dans l'image et sa position dans les images précédentes, nous cherchons à déterminer la nature de l'objet détecté. Un historique de l'occupation de la pièce est ensuite sauvegardé dans un fichier texte.
Les performances de cet algorithme sont mises en évidence au travers de quelques résultats expérimentaux. Enfin, nous présentons les conclusions et perspectives de ce travail.

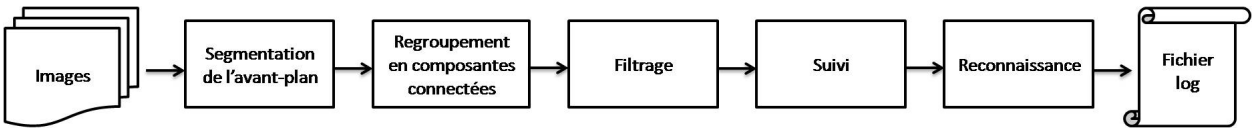

Fig. 1. Processus mis en place

1 Segmentation de l'avant-plan

La première étape de l'algorithme consiste en la segmentation l'avant-plan. Celle-ci est une étape primordiale car les étapes suivantes sont fortement dépendantes de la qualité de cette segmentation. L'objectif de ce traitement est de simplifier au maximum l'image, sans altérer les informations, pour ne laisser aux étapes suivantes que quelques régions d'intérêt (régions de l'image où il y a une forte probabilité qu'il y ait un

homme). Par définition, l'arrière-plan est l'union de tous les objets statiques correspondant à la scène et l'avant-plan représente tous les objets susceptibles d'être des humains. Il existe deux grandes catégories de méthodes : les algorithmes basés sur la différence de deux (ou trois) images successives, et les algorithmes de soustraction de l'arrière-plan. Les modèles de l'arrière-plan les plus utilisés sont un moyennage temporel, une distribution gaussienne [7, 11], un mélange de gaussiennes [9] ou un minimum et un maximum [8].
Nous avons montré dans [10], qu'une soustraction de l'arrière-plan avec une modélisation par une distribution gaussienne présente de bonnes performances en terme de détection et en terme de temps de calcul. De plus, comme nous ne nous intéressons qu'à des environnements intérieurs, il est inutile d'utiliser un modèle multi-modal, du type mélange de gaussiennes. Nous avons donc choisi de modéliser chaque pixel de l'arrière-plan par une distribution gaussienne. On calcule donc tout d'abord la valeur moyenne et la variance de chaque pixel de l'arrière-plan. Puis, la détection de l'avant-plan se fait par :

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

où $I_t(x,y)$ représente la valeur du pixel de coordonnées (x,y) à l'instant t ; $B_{1,t}$ est l'image binaire représentant l'avant-plan détecté ; μ_t représente la valeur moyenne ; σ_t l'écart type et τ_1 est un seuil fixé empiriquement à 2.5.
Le modèle gaussien est mis à jour si $B_{1,t}(x,y) = 0$. Soit :

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

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

où α est un seuil déterminé empiriquement.
En plus de permettre une vision nocturne, une caméra infrarouge apporte une information sur la température des éléments dans son cône de vision. En partant du principe qu'un humain est sensiblement plus chaud que son environnement, nous effectuons une binarisation de l'image pour mettre en relief les zones chaudes de l'image.

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

où $B_{2,t}$ représente l'image binaire des zones chaudes, $I(x,y)$ la valeur en niveau de gris du pixel de coordonnées (x,y), τ_2 un second seuil déterminé empiriquement. Nous effectuons ensuite un simple "et logique" entre l'image binaire de la soustraction de l'arrière-plan et l'image binaire des zones chaudes.

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

où B_t est le résultat de notre segmentation de l'avant-plan. Nous regroupons ensuite les composantes connectées et nous supprimons les composantes trop petites. Un exemple de segmentation de l'avant-plan, après filtrage, est présenté figure 2.

Fig. 2. Exemple de segmentation de l'avant-plan

2 Suivi de cibles

Après avoir détecté les régions d'intérêt dans l'image, nous souhaitons avoir un historique de leurs déplacements dans le plan image. Pour conserver les performances "temps-réel", nous avons développé un algorithme de suivi relativement simple et rapide, basé sur le recouvrement des composantes connectées entre les trames successives. Nous cherchons donc la correspondance entre les composantes de l'image à l'instant t avec les composantes à l'instant $t-1$. Pour cela, nous calculons H_t, la matrice de correspondance à l'instant t :

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

où M et N correspondent respectivement aux nombres de composantes connectées à l'instant $t-1$ et à l'instant t. $\beta_{i,j} = 1$ si la composante i à l'instant $t-1$ et la composante j à l'instant t se recouvre, $\beta_{i,j} = 0$ sinon. L'analyse de la matrice H_t nous permet de connaître la correspondance entre les composantes de l'image à l'instant t avec les composantes à l'instant $t-1$. Par exemple, si deux composantes a et b à l'instant $t-1$ et une à l'instant t se recouvrent, nous fusionnons les deux composantes a et b en une seule composante. Notre algorithme est capable de gérer les regroupements entre plusieurs composantes et la séparation d'une composante en plusieurs. Cependant, comme nous n'utilisons aucun modèle pour la cible suivie et que nous n'estimons pas le mouvement de nos cibles, nous ne sommes pas capables de gérer les occultations. Pour notre application, cela n'a pas beaucoup de conséquence : si un objet disparaît, il sera considéré comme étant un nouvel objet lorsqu'il réapparaîtra.

3 Reconnaissance d'un humain

Une fois la région d'intérêt détectée et suivie, nous souhaitons connaître la nature de l'objet, en l'occurence, si c'est un humain. La région d'intérêt détectée précédemment est donc analysée.
Pour cela, il est possible d'extraire certaines caractéristiques de la région d'intérêt (contours, couleurs, textures...) pour trouver une combinaison de ces caractéristiques spécifiques à notre classe (les humains). Mais pour des objets complexes, il est très difficile de trouver un modèle générique. Les humains ont différentes tailles, couleurs, le corps est articulé ... C'est pourquoi nous avons préféré prendre le parti de construire un modèle statistique par les techniques d'apprentissage.
Nous avons donc besoin d'une base d'apprentissage composée d'exemples positifs et négatifs (images qui contiennent ou non un humain). Durant l'apprentissage, différentes caractéristiques sont extraites des exemples positifs et négatifs et un modèle statistique est construit. Il existe dans la littérature beaucoup de descripteurs et beaucoup de techniques d'apprentissages. Nous avons choisi d'utiliser le système initialement proposé par Viola et Jones [5] pour détecter des visages. Cette méthode est basée sur les ondelettes de Haar et l'algorithme du boosting Adaboost.
Notre base d'apprentissage est composé de 3965 images négatives et 956 images positives (cf. figure 3). Les images viennent des bases d'images OTCBVS [12, 13] et d'images collectées avec une caméra infrarouge dont la vérité terrain a été manuellement définie.

Fig. 3. Exemples d'images positives

Nous utilisons l'ensemble des 14 descripteurs décrits figure 4. Chaque descripteur est composé de deux ou trois rectangles blancs et noirs. La valeur du descripteur x_i est calculée par une somme pondérée de la valeur des pixels de chaque composante noire et blanche.
Chaque descripteur est ensuite utilisé comme un classifieur faible, tel que :

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

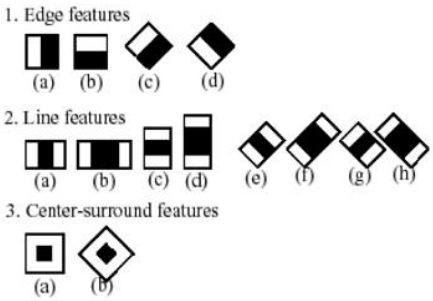

Fig. 4. Ensemble des descripteurs utilisés

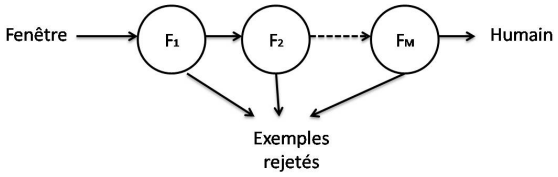

Fig. 5. Cascade de classifieurs boostés

où $+1$ correspond à la présence d'un homme dans la fenêtre d'entrée et -1 non, τ_i est un seuil. Un classifieur plus robuste est ensuite construit avec plusieurs classifieurs faibles par la méthode du boosting [14].

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

Ensuite, une cascade de classifieurs boostés est construites (cf. figure 5). Une fenêtre d'entrée est analysée successivement par chaque classifieur boosté F_k qui peut envoyer la fenêtre au classifieur suivant ou rejeter la fenêtre. Les classifieurs simples sont placés en premier, ils permettent de rejeter rapidement un grand nombre de fenêtres correspondant à l'arrière-plan.

4 Résultats expérimentaux

La vitesse de notre détecteur est étroitement liée au nombre et à la taille des régions d'intérêt. Cependant, pour une vidéo de taille $564 * 360$, notre algorithme est capable de traiter approximativement 30 images par secondes lorsqu'il n'y a aucune région d'intérêt à analyser et de 15 à 20 images par seconde lorsqu'il y a une région d'intérêt à analyser. Cette vitesse d'execution est compatible avec les contraintes temps réel de notre système. Un exemple de détection est montré figure 6. L'ellipse verte correspond à la région d'intérêt détectée, un rectangle rouge s'affiche ensuite s'il y a un humain au voisinage de cette région d'intérêt.

Fig. 6. Exemple de détection

Conclusion

Nous avons présenté dans cet article un système complet de détection de la présence humaine dans un environnement intérieur basé sur le système de détection de visages [5]. Ce système permet une réduction de l'espace de recherche dans chaque trame en recherchant des objets chauds en mouvement dans la vidéo. Avec un module de suivi de cibles, nous sommes capables d'avoir un historique des déplacements dans la pièce. Cet historique sera ensuite utilisé dans des travaux ultérieurs pour valider des algorithmes de vision avec du matériel à moindre coût (dans le domaine spectral visible ou proche infrarouge) ou d' autres technologies (ultrason, capteur pyroélectrique ...).
Les résultats expérimentaux ont montré les performances de notre approche. Cependant, il existe encore de nombreux axes de travail. Tout d'abord, nous devons travailler sur la qualité de la base d'images pour l'apprentissage, la performance du classifieur est étroitement liée avec la qualité de cette base d'images. Nous devons également apprendre plusieurs classifieurs pour un humain. En effet, comme nous travaillons en environnement intérieur, les occultations sont fréquentes, il serait donc judicieux d'apprendre, en plus du corps entier, une autre partie du corps qui est plus souvent visible (e.g. la tête et les épaules).

Références

1. Y. Kuno, T. Watanabe, Y. Shimosakoda and S. Nakagawa, "Automated Detection of Human for Visual Surveillance System", Proceedings of the International Conference on Pattern Recognition, 865–869, 1996
2. Y. Dedeoglu, "Moving object detection, tracking and classification for smart video surveillance", PhD thesis, bilkent university, 2004
3. Y. Mae, N. Sasao, K. Inoue T. Arai, "Person detection by mobile-manipulator for monitoring", The Society of Instrument and Control Engineers Annual Conference, 2003
4. C. Papageorgiou, M. Oren and T. Poggio, "A general framework for object detection", 6th International Conference on Computer Vision, 555–562, 1998
5. P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features", Proceedings of the conference on Computer Vision and Pattern Recognition, 511–518, 2001
6. N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection", IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 886–893, 2005
7. S. Yoon and H. Kim, "Real-time multiple people detection using skin color, motion and appearance information", Proceedings of the International Workshop on Robot and Human Interactive Communication, 331-334, 2004
8. I. Haritaoglu, D. Harwood, and LS. David. "W4 : real-time surveillance of people and their activities", IEEE Transaction on Pattern Analysis and Machine Intelligence, 809-830, 2006
9. C. Stauffer and E. Grimson, "Adaptive background mixture models for real-time tracking", Proceedings of the conference on Computer Vision and Pattern Recognition, 246–252, 1999
10. Y. Benezeth, B. Emile and C. Rosenberger, "Comparative Study on Foreground Detection Algorithms for Human Detection", Proceedings of the Fourth International Conference on Image and Graphics, 661–666, 2007
11. J. Han and B. Bhanu, "Detecting moving humans using color and infrared video", Proceedings of IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems, 228–233, 2003
12. J. Davis and M. Keck, "A two-stage approach to person detection in thermal imagery", In Proceedings Workshop on Applications of Computer Vision, 2005
13. J. Davis and V. Sharma, "Background-Subtraction using Contour-based Fusion of Thermal and Visible Imagery", Computer Vision and Image Understanding, 162–182, 2007
14. R.E. Schapire, "The boosting approach to machine learning: An overview", MSRI Workshop on Nonlinear Estimation and Classification, 2002