Détection de la présence humaine et caractérisation de l'activité *

Yannick BENEZETH[1], Bruno EMILE[2], Hélène LAURENT[1], Christophe ROSENBERGER[3]

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

[2]Institut PRISME, Université d'Orléans
2 av. F. Mitterand, 36000 Châteauroux, France

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

yannick.benezeth@ensi-bourges.fr, Bruno.Emile@univ-orleans.fr,
helene.laurent@ensi-bourges.fr, Christophe.Rosenberger@greyc.ensicaen.fr

Résumé – Nous présentons dans cet article une méthode de détection de personnes et une mesure de leur activité dans des séquences vidéos. En posant l'hypothèse que notre caméra est fixe, nous utilisons tout d'abord la soustraction de l'arrière-plan pour réduire l'espace de recherche de notre classifieur. Nous utilisons ensuite le suivi de points d'intérêt et l'analyse des composantes connectées détectées par la soustraction de l'arrière-plan pour suivre les objets dans le plan image. La classification est réalisée grâce à plusieurs cascades de classifieurs boostés. Cette méthode a été évaluée sur une large base de vidéos. Nous présentons enfin une méthode pour caractériser l'activité des personnes présentes dans une pièce.

Abstract – In this paper, we present a human detection system for the analysis of videos sequences. We perform first background subtraction in order to reduce the search space of the classifier. A tracking step based on connected components analysis combined with feature points tracking allows to collect information on 2D movements of objects in the image plane and so to improve the performance of our classifier. A classification based on four cascades of boosted classifiers is used for the recognition. This method has been evaluated over a wide range of real-life videos. Then, we present a method for caracterizing people activity in a room.

1 Introduction

Il est facile pour un humain de reconnaître un autre homme autour de lui ou sur une image mais c'est un problème très complexe pour un système automatisé. Pourtant, beaucoup de systèmes ont besoin d'avoir des informations sur la présence ou l'absence de personnes dans leur environnement. Les applications potentielles d'un système de détection automatique de la présence humaine sont nombreuses. Nous pouvons par exemple citer les systèmes de transport intelligent, la robotique, la surveillance, la domotique intelligente, l'indexation d'images ou de vidéos ... Nos travaux s'inscrivent dans le cadre du projet CAPTHOM du pôle de compétitivité $S2E2$ qui vise à développer un capteur intelligent pour détecter la présence humaine afin de réguler la consommation d'énergie dans le bâtiment (éclairage, chauffage *etc.*). Pour ce faire, nous présentons dans cet article une méthode permettant de récupérer des informations sur la présence, le nombre et l'activité des personnes dans une pièce.

D'une manière générale, le principe de la détection de personnes dans une image ou une vidéo est le même. Cependant, pour une image, le détecteur ne dispose d'aucune information *a priori* et doit donc parcourir l'image entière avec une fenêtre de détection. Pour une vidéo, il est possible de simplifier le problème en se focalisant par exemple sur les zones de l'image où il y a eu un mouvement. Il est également possible de raisonner en fonction de la position des personnes présentes dans les images précédentes ou même de caractériser les personnes par leur mouvement.

On trouve principalement deux groupes de méthodes dans la littérature pour détecter un humain dans une image. Il y a d'une part les méthodes qui utilisent un modèle explicite (2D ou 3D) de la forme du corps humain (*e.g.* [1, 2]) et d'autre part celles qui se basent sur des techniques d'apprentissage supervisé. À partir d'une base d'images, des caractéristiques de la forme du corps humain sont extraites et un modèle discriminant est construit. Nous pouvons citer par exemple Papageorgiou *et al.* [3] qui ont proposé un détecteur basé sur les ondelettes de Haar et les machines à vecteurs de support. Viola et Jones [4, 7] ont également proposé un détecteur basé sur les filtres de Haar et l'algorithme du boosting. Plus récemment, Dalal et Triggs [5] ont utilisé avec succès les histogrammes de gradients orientés dans le cas de la détection de piétons. D'autres méthodes détectent séparément différentes parties du corps humain et fu-

*Ces travaux ont été réalisés grâce au support financier du conseil régional de la région Centre, du ministère de l'industrie dans le cadre du projet CAPTHOM du pôle de compétitivité $S2E2$

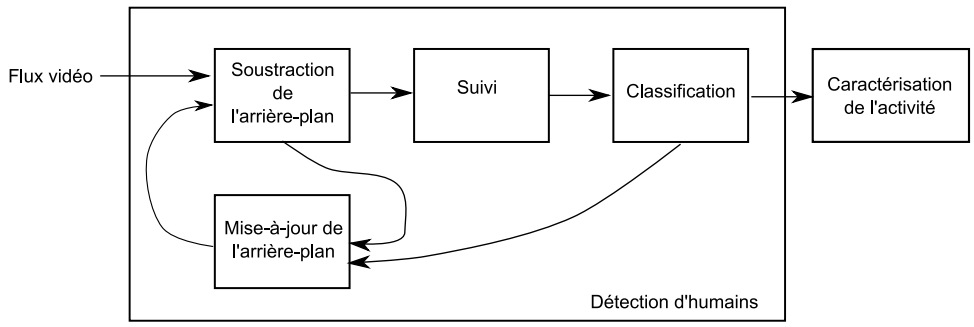

FIGURE 1: Processus mis en place pour la détection de personnes et la caractérisation de l'activité dans des séquences vidéos.

sionnent ensuite ces résultats [6].

Dans cet article, nous présentons un système de détection de personnes où l'espace de recherche du classifieur est réduit grâce à une soustraction de l'arrière-plan. Un suivi des objets est basé sur l'analyse des composantes connectées et le suivi de points d'intérêt. Cet historique du déplacement dans le plan image nous permet d'augmenter grandement les performances du système global. La classification est réalisée avec plusieurs cascades de classifieurs boostés utilisant l'algorithme adaboost et les ondelettes de Haar [4]. Nous utilisons précisément quatre classifieurs différents. Finalement, une méthode de caractérisation de l'activité est présentée. La figure 1 schématise ce processus. Ce système a été évalué sur une large base de vidéos. Dans la suite de cet article, nous présentons la méthode de détection de personnes puis la mesure de l'activité proposée. Ensuite, nous présentons les résultats expérimentaux obtenus sur une large base de vidéos.

2 Détection d'humains

Dans cette partie, la méthode de détection de personnes dans des séquences vidéos est présentée. Celle-ci est composée de trois modules différents, décrits ci-dessous.

2.1 Soustraction de l'arrière-plan

Puisque notre caméra est statique, l'espace de recherche peut être réduit en détectant les zones d'intérêt grâce à la soustraction de l'arrière plan. Suite à une étude comparative des méthodes de soustraction de l'arrière-plan [8], nous avons choisi de modéliser chaque pixel de l'arrière-plan avec une distribution gaussienne. Malgré la simplicité de ce modèle, les performances obtenues sont très largement satisfaisantes lorsqu'il est utilisé dans un environnement sans difficulté particulière. De plus, ce modèle présente l'avantage de demander peu de ressources en termes de calcul et peu d'espace mémoire. La distance entre les pixels de l'image courante et le modèle de l'arrière-plan est calculée avec la distance de Mahalanobis puis seuillée pour détecter les changements :

$$\mathcal{X}_t(s) = \begin{cases} 1 & \text{si } |\boldsymbol{I}_{s,t} - \boldsymbol{\mu}_{s,t}|\boldsymbol{\Sigma}_{s,t}^{-1}|\boldsymbol{I}_{s,t} - \boldsymbol{\mu}_{s,t}|^T > \tau_1 \\ 0 & \text{sinon,} \end{cases} \quad (1)$$

où $\boldsymbol{I}_{s,t}$ représente la distribution de couleur du pixel s à l'instant t, $\boldsymbol{\mu}$ la moyenne, $\boldsymbol{\Sigma}$ la matrice de covariance, τ_1 un seuil déterminé empiriquement et $\mathcal{X}_t$ l'image de l'avant-plan. Le modèle de l'arrière-plan est mis à jour à chaque itération à trois niveaux différents. Tout d'abord, chaque pixel est mis à jour avec un filtre moyenneur temporel. Ensuite, si le classifieur nous donne, avec un niveau de confiance suffisant, l'information qu'un nouvel objet statique ajouté à la scène n'est pas un humain, tous les pixels de l'avant-plan correspondants seront inclus dans le modèle de l'arrière-plan pour éviter d'effectuer de nouveaux traitements inutiles (classification et suivi). Finalement, si un changement d'illumination globale est détecté, le modèle de l'arrière-plan est réinitialisé.

2.2 Suivi

Après avoir filtré l'image de l'avant-plan détecté, nous regroupons en composantes connectées les pixels de l'avant-plan. À chaque instant, nous disposons donc de la liste des composantes connectées présentes et de la liste des objets suivis dans les images précédentes. Nous cherchons ensuite à faire la correspondance entre ces deux listes. Nous utilisons pour cela la matrice de correspondance $\mathcal{H}_t$ définie par :

$$\mathcal{H}_t = \begin{pmatrix} \beta_{1,1} & \cdots & \beta_{1,N} \\ \vdots & \ddots & \vdots \\ \beta_{M,1} & \cdots & \beta_{M,N} \end{pmatrix} \quad (2)$$

où M correspond au nombre d'objets suivis et N au nombre de composantes connectées présentes sur l'image de l'avant-plan à l'instant t. $\beta_{i,j} = 1$ s'il y a correspondance entre l'objet suivi i et la composante connectée j, sinon $\beta_{i,j} = 0$. Chaque objet suivi est caractérisé par un ensemble de points d'intérêt. Ces points sont suivis, image par image, et la position de ces points par rapport aux composantes connectées permet de faire la correspondance entre les objets suivis et les composantes connectées détectées. Le suivi des points d'intérêt est réalisé avec la méthode de Lucas et Kanade [9, 10]. Deux contraintes sont ajoutées à la méthode originale :

1. un point suivi doit être sur un pixel de l'avant-plan. Dans le cas contraire, le point est supprimé de la liste de points et un nouveau est créé.
2. Lors de la création d'un nouveau point, une contrainte de distance avec les autres points est imposée de manière à avoir une répartition homogène des points sur tout l'objet.

La correspondance entre les objets suivis et les composantes connectées est réalisée en calculant à quel objet appartiennent les points présents sur une composante connectée. Soit $\gamma_{i,j}$ le nombre de points appartenant à l'objet i présents sur la composante j.

$$\begin{cases} \beta_{i,j} = 1 & \text{si } \gamma_{i,j} > \tau_2 \\ \beta_{i,j} = 0 & \text{sinon.} \end{cases} \quad (3)$$

Le seuil τ_2 dépend directement du nombre de points utilisés pour représenter un objet. En pratique, le seuil τ_2 est fixé à 25% du nombre de points d'intérêt par objet. À partir de cette matrice de correspondance, nous sommes capable d'établir un ensemble de règles pragmatiques afin de gérer les cas les plus courants. Nous gérons les 5 cas suivants :

- la simple correspondance,
- la séparation d'un objet en plusieurs,
- la fusion de plusieurs objets en un seul,
- la suppression d'un objet,
- la création d'un nouvel objet.

2.3 Classification

Une fois que nous avons établi l'historique du déplacement des objets dans le plan image, nous souhaitons connaître la nature de ces objets, en l'occurrence, si nous suivons des humains. Cette capacité de ne détecter que les humains et non plus tous les objets mobiles est un élément clé pour notre système et constitue une différence très importante avec les détecteurs de présence actuellement sur le marché [11]. Nous utilisons 4 détecteurs au voisinage de l'objet suivi, à savoir :

- un détecteur des personnes debout, quel que soit le point de vue (de face, de profil etc.),
- un classifieur de la partie supérieure du corps, vue de face ou de dos,
- un classifieur de la partie supérieure du corps, vue de gauche,
- et finalement un classifieur de la partie supérieure du corps, vue de droite,

Chaque détecteur est construit avec la méthode proposée initialement par Viola *et al.* [4] : un détecteur est composé d'une cascade de classifieurs boostés, construite avec l'algorithme adaboost. Nous utilisons précisément 14 filtres, illustrés dans la figure 2. Ces filtres sont composés de 2 ou 3 rectangles blancs et noirs. La valeur du descripteur x_i associé au filtre i est calculée par une somme pondérée des pixels de chaque partie du filtre.

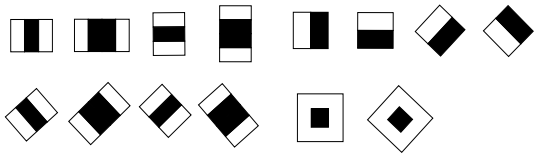

FIGURE 2: Illustration des filtres de Haar utilisés

À partir de chaque descripteur x_i, on obtient ensuite un classifieur faible f_i par l'expression suivante :

$$f_i = \begin{cases} +1 \text{ si } x_i \geq \tau_i \\ -1 \text{ si } x_i < \tau_i \end{cases} \quad (4)$$

où $+1$ correspond à la présence d'une personne et -1 non. Le seuil τ_i correspond au seuil optimal qui minimise le taux de mauvaise classification du classifieur faible sur la base d'apprentissage. Ensuite, un classifieur plus robuste est construit avec plusieurs classifieurs faibles, pondérés par les poids c_i, avec la méthode du boosting [12] :

$$F_j = sign(c_1 f_1 + c_2 f_2 + \ldots + c_n f_n). \quad (5)$$

Une cascade de classifieurs boostés est finalement construite (cf. figure 3). F_j correspond au classifieur boosté du j^{eme} étage de la cascade. Chaque étage peut rejeter ou accepter une fenêtre d'entrée, si une fenêtre passe tous les étages de la cascade, l'algorithme labellise comme humain la fenêtre analysée. Cette méthode a été choisie pour ses bonnes performances de détection et aussi car l'utilisation des images intégrales, sur laquelle elle se base, permet de calculer les filtres de Haar très rapidement et de répondre aux contraintes temps réel de notre application.

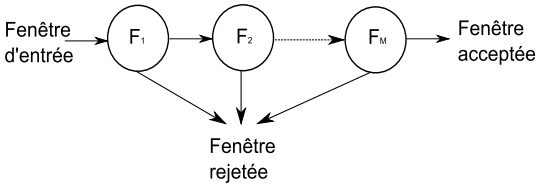

FIGURE 3: Cascade de classifieurs

Finalement, nous combinons les résultats des 4 détecteurs par parties pour construire un indice de confiance de l'appartenance à la classe humain de l'objet suivi. Cet indice évoluera dans le temps selon les résultats de détection à chaque instant. Un simple seuillage de cet indice de confiance nous informe sur la nature de l'objet suivi. La figure 4 illustre les différentes étapes mises en oeuvre pour la détection.

3 Caractérisation de l'activité

Connaissant l'historique des positions de chaque personne présente dans la pièce surveillée, nous souhaitons caractériser leur activité. Cette information est très importante car elle permet d'estimer l'apport calorifique nécessaire, compte tenu de l'activité des personnes présentes dans une pièce, pour obtenir la température ambiante conduisant au confort optimal. Nous proposons une démarche basée sur deux approches pour obtenir une mesure de l'activité qui sera proportionnelle à l'agitation et au déplacement de la personne. Tout d'abord, nous proposons d'analyser les chemins parcourus par les occupants et leur vitesse dans le plan image. Dans un deuxième temps, nous prenons en compte les normes des vecteurs déplacements des points d'intérêt utilisés lors du suivi de cibles pour être capable de détecter une personne "agitée" mais qui ne se déplace

FIGURE 4: Illustration des différentes étapes mises en oeuvre pour la détection. De gauche à droite : image d'entrée, soustraction de l'arrière-plan, suivi et résultat final avec affichage de l'indice de confiance.

pas. Une combinaison de ces deux indices permet ultérieurement d'accéder très simplement à une caractérisation de l'activité des personnes présentes dans une pièce.

4 Résultats expérimentaux

Le système de détection de personnes présenté ci-dessus a été évalué sur une large base de 29 vidéos représentant plusieurs situations (réunion de travail, zone de passage *etc.*) dans plusieurs environnements (laboratoire, salle à manger *etc.*). Le contenu de chaque vidéo a été manuellement annoté. Nous présentons ci-dessous le résultat obtenu, sous la forme d'une matrice de confusion, de la réponse du système à la question suivante : est-ce que quelqu'un est présent dans le champ de vision de la caméra ? La méthode proposée ci-dessus a été comparée avec le système proposé par Viola *et al.* [4].

	+	-
+	0.77	0.23
-	0.58	0.42

(a)

	+	-
+	0.97	0.03
-	0.03	0.97

(b)

TABLE 1: Matrices de confusion obtenues sur une base de 29 vidéos par l'algorithme de Viola *et al.* [4] et par l'algorithme proposé (b). Les lignes correspondent à la vérité terrain tandis que les colonnes correspondent au résultat de l'algorithme, $+$ signifie présence et $-$ absence.

Les résultats présentés ci-dessus montre des performances très satisfaisantes puisqu'avec un taux de détection d'environ 97%, il n'y a que 3% de fausses détections. La réduction de l'espace de recherche du classifieur avec une soustraction de l'arrière-plan nous permet de diminuer le nombre de fausses détections et le suivi nous permet de diminuer le nombre de détections manquées.

5 Conclusion

Nous avons présenté dans cet article un système de détection de personnes dans un environnement intérieur. Une soustraction de l'arrière-plan permet de limiter l'espace de recherche du classifieur et un suivi des objets détectés permet d'augmenter la fiabilité de notre classifieur. Nous utilisons une combinaison de quatre détecteurs pour effectuer la classification. Dans un second temps, nous avons présenté une mesure permettant de caractériser l'activité des personnes présentes dans une pièce. Ces informations seront utiles au système de gestion technique de bâtiment pour réguler la consommation d'énergie électrique (éclairage, chauffage *etc.*). Ce système a été évalué sur une large base de vidéos. Les résultats expérimentaux présentés ici sont très encourageants. Par la suite, nous nous servirons du résultat de ce système, c'est-à-dire la position des personnes dans chaque image, pour étendre les possibilités de notre capteur. Nous prévoyons notamment d'ajouter des fonctionnalités d'identification, de reconnaissance de postures *etc.*

Références

[1] Q. Zhao, J. Kang, H. Tao and W. Hua, "Part Based Human Tracking In A Multiple Cues Fusion Framework", *in Proc. of the International Conference on Pattern Recognition*, 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, M. Oren and T. Poggio, "A general framework for object detection", *in Proc. of the IEEE International Conference on Computer Vision*, pp. 555–562, 1999.

[4] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features", *in IEEE Proc. of the conference on Computer Vision and Pattern Recognition*, pp. 511–518, 2001.

[5] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection", *in IEEE Proc. of the conference on Computer Vision and Pattern Recognition*, pp. 886–893, 2005.

[6] B. Wu and R. Nevatia, "Tracking of multiple, partially occluded humans based on static body part detection", *in IEEE Proc. of the conference on Computer Vision and Pattern Recognition*, pp. 951–958, 2006.

[7] P. Viola, M.J. Jones and D. Snow, "Detecting pedestrians using patterns of motion and appearance", *in International Journal of Computer Vision*, pp. 153–161, 2005.

[8] Y. Benezeth, P.M. Jodoin, B. Emile, H. Laurent, C. Rosenberger, "Review and evaluation of commonly-implemented background subtraction algorithms", *in Proc. of the International Conference on Pattern Recognition*, 2008.

[9] B. Lucas, T. Kanade, "An iterative image registration technique with an application to stereo vision", *in Proc. of the International Joint Conference on Artificial Intelligence*, pp 674-679, 1981.

[10] J. Shi and C. Tomasi, "Good features to track", *in Proc. of the international conference on Computer Vision and Pattern Recognition*, pp. 593-600, 1994.

[11] J.F. Gobeau, "Détecteurs de mouvement à infrarouge passif (détecteurs IRP)", *Colloque Capteurs*, 2007.

[12] R.E. Schapire, "The boosting approach to machine learning : An overview," *in Workshop on N.E.C.*, 2002.