Analyse temps réel des postures humaines dans une foule avec des images infrarouges.

Lætitia GOND, Quoc-Cuong PHAM, Julien BEGARD, Nicolas ALLEZARD, Patrick SAYD

CEA, LIST,

Laboratoire Systèmes de Vision Embarqués

Boîte Courrier 65, Gif-sur-Yvette, F-91191 France

{prenom.nom}@cea.fr

Résumé – Cet article présente un système de vidéosurveillance développé dans le cadre du projet ISCAPS. L'imagerie infrarouge lointain représente un moyen robuste de gérer les changements de visibilité pouvant intervenir lors des acquisitions (luminosité, fumée), et de distinguer plus facilement des humains dans des scènes complexes. Dans cet article, nous démontrons en particulier son efficacité pour l'analyse des postures dans un groupe compact de personnes. Notre objectif est de détecter automatiquement la chute de plusieurs personnes dans une foule dense. La méthode présentée ici est basée sur la détection et la segmentation d'individus dans le groupe de personnes, grâce à l'utilisation d'une combinaison de plusieurs classifieurs faibles. L'analyse des silhouettes ainsi extraites permet de détecter les situations anormales. Notre approche a été appliquée avec succès au contexte de détection d'attaques chimiques sur un quai de gare et validée expérimentalement dans le projet. Des résultats expérimentaux sont présentés dans cet article.

Abstract – This article describes a video-surveillance system developed within the ISCAPS project. Thermal imaging provides a robust solution to visibility change (illumination, smoke) and is a relevant technology for discriminating humans in complex scenes. In this article, we demonstrate its efficiency for posture analysis in dense groups of people. The objective is to automatically detect several persons lying down in a very crowded area. The presented method is based on the detection and segmentation of individuals within groups of people using a combination of several weak classifiers classifiers. The classification of extracted silhouettes enables to detect abnormal situations. This approach was successfully applied to the detection of terrorist gas attacks on railway platform and experimentally validated in the project. Some of the results are presented here.

1. Introduction

L'objectif général du projet ISCAPS est de proposer des réponses technologiques au risque d'attaques terroristes dans les lieux publics afin de les éviter ou de limiter leurs conséquences. Ces réponses passent par la mise en place de moyens de surveillance à la fois simples d'utilisation et efficaces, fonctionnant de façon automatique et en temps réel. Cet article décrit l'un des systèmes développés dans ce projet, répondant aux spécifications d'un opérateur de transport public (SNCF). Le scénario est le suivant : sur le quai d'une gare, le système doit détecter au plus tôt une éventuelle attaque chimique, simplement grâce à l'analyse du comportement des personnes présentes dans la scène. Il semble en effet clair que dans un pareil cas, une détection précoce peut permettre de réduire considérablement les dégâts engendrés [15]. Deux cas de figures importants se distinguent. Premièrement, à la suite d'un incendie ou d'une attaque au gaz, la zone peut se retrouver complètement enfumée, engendrant probablement un mouvement de panique collective. La plupart des personnes parviendront à s'enfuir, mais d'autres se retrouveront bloquées dans un environnement particulièrement dangereux. Deuxièmement, en l'absence de détection des gaz, un incident peut tout de même être détecté par l'observation des réactions des personnes présentes, comme le vacillement, des quintes de toux ou des chutes. Parmi les contraintes du scénario figurent la capacité du système à fonctionner malgré la fumée ou l'obscurité, et l'interprétation automatique du comportement de personnes en cas de malaise, qui représentent deux tâches difficiles. Un capteur infrarouge non-refroidi (technologie micro-bolomètre, 8-12 μm) est utilisé pour apporter de la robustesse vis-à-vis des conditions de visibilité difficiles (FIG.1). Notre système doit être à la fois capable d'estimer le nombre de personnes présentes dans un lieu envahi par la fumée, mais aussi de détecter d'éventuelles chutes. Jusqu'à présent, en raison de leur coût et de leur faible durée de vie, l'utilisation de capteurs infrarouges restait confinée au domaine militaire, pour des applications comme la détection ou le suivi de véhicules. Avec l'apparition de la nouvelle génération de capteurs IR non-refroidis, moins coûteux et plus robustes,

un nouveau champ d'application s'est ouvert. Leurs bonnes performances dans des conditions de visibilité difficiles et leur faculté à détecter aisément des personnes (grâce à l'émission IR naturelle des êtres vivants), en font un outil prometteur pour des applications comme la surveillance de sites [4] ou l'assistance à la conduite [14, 17]. Dans le cas d'ISCAPS, c'est la robustesse de ces capteurs vis-à-vis de la présence de fumée qui nous a conduit à choisir cette technologie. Les images infrarouges sont toutefois monochromatiques, et leur texture reste relativement pauvre si on les compare à celle des images du spectre visible.

FIG. 1: Influence de la fumée dans des images couleur et infrarouge. Haut sans fumée, bas avec fumée

Dans notre cas, l'analyse de la scène est rendue complexe d'une part par la densité de la foule faisant face à la caméra et d'autre part par la présence possible de bagages sur le quai de la gare.
Indépendamment de la technologie d'imagerie utilisée, la plupart des techniques de vidéosurveillance se sont concentrées sur l'analyse de scènes dans lesquelles les personnes ne se chevauchent pas ou peu dans les images. La complexité et la variabilité des scènes de foule (nombre de personnes, occultations, postures) requièrent l'utilisation d'outils bien spécifiques. Concernant la détection,[19] propose une méthode capable de gérer le cas de personnes partiellement occultées, mais ne considère pas réellement le problème des groupes très denses. Dans [6], une méthode basée sur le flot optique et l'extraction de contours permet d'estimer la densité et le mouvement d'une foule. Cette approche n'est toutefois pas fiable dans le cas de foules denses. Les auteurs de [12] parviennent à effectuer le suivi global d'un groupe et à évaluer sa densité. Dans [13], la densité d'une foule est estimée grâce au calcul de la dimension fractale des contours. [1] présente un détecteur de situations d'urgence dans des foules, s'appuyant sur des statistiques du flot optique extraites de données vidéo. Toutefois, aucune de ces approches n'a été conçue pour détecter le comportement anormal de certains individus dans la foule. Dans notre application, cette analyse représente pourtant une étape essentielle. Certaines approches ont abordé ce sujet. Dans [20], un algorithme de segmentation bayésien a été proposé pour dénombrer les personnes dans la foule grâce à des modèles de forme. Mais cette méthode est trop lente dans le cas d'une grande foule. [16] présente un algorithme de détection dans une foule basé sur l'analyse spatio-temporelle d'une séquence vidéo. Une segmentation des régions en mouvement est combinée avec une classification des piétons, de la foule et des véhicules. Cette approche est intéressante pour compter les personnes plus que pour extraire certains individus et elle ne permet pas d'analyser des personnes statiques. Notre approche consiste à traiter les images infrarouges dans le but d'extraire des individus d'un groupe dense et propose une solution innovante pour détecter des personnes à terre.

2. Résumé de la méthode proposée

La détection des personnes à terre est rendue difficile par la grande variabilité de leurs apparences (allongées, roulées en boule...). Notre méthode consiste à extraire les objets d'intérêt de la scène. Parmi ces objets, les personnes debout sont segmentées et supprimées afin de permettre une meilleure analyse des objets restant et d'en extraire les personnes couchées ou agenouillées.
Comme notre caméra IR est fixe, nous avons choisi d'effectuer une modélisation du fond pour extraire les objets d'intérêt dans un module de prétraitement. Le fond est donc appris grâce à une méthode statistique adaptative. En raison de la grande variabilité des vêtements et des postures, et de la densité potentiellement forte des personnes présentes dans l'image, la forme et l'apparence de la tête nous semble être une caractéristique visuelle plus stable que l'individu tout entier. Dans la deuxième phase de notre algorithme, les hypothèses sur la présence d'individus sont donc générées par un détecteur de têtes, qui combine trois techniques complémentaires i) une détection des pics sur l'ensemble des blobs obtenus après soustraction de fond, ii) une détection des formes elliptiques dans l'image infrarouge, et iii) une détection de la forme représentée par l'ensemble tête-épaules.

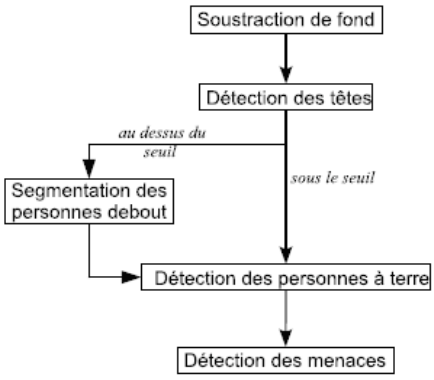

FIG. 2 : Les différentes étapes de l'algorithme.

Ces hypothèses de détection sont ensuite classées en deux groupes, en fonction de leur position dans la scène par rapport à un seuil sur la hauteur : les têtes situées au dessus du seuil permettent d'initialiser un modèle de personne debout. Les paramètres de ce modèle sont ajustés lors d'une étape de segmentation, grâce à une méthode de

Monte Carlo par chaînes de Markov (MCMC). Ce raffinement permet d'obtenir une meilleure localisation des personnes debout, pour améliorer par la suite l'analyse des composantes restantes. Les blobs restants sont examinés et classés comme étant une personne couchée ou un autre objet. Les têtes détectées situées en dessous du seuil de hauteur constituent quant à elles des hypothèses de personnes couchées. Ces données sont enfin fournies en entrée d'un module de détection de menace, qui estime le risque de manière probabiliste, et déclenche différents niveaux d'alarme. Une vue d'ensemble de notre algorithme est présentée sur la figure 2.

3. Segmentation des humains dans des images infrarouges

3.1 Modélisation du fond

Dans les images thermiques, les personnes se distinguent plus facilement du fond que dans les images couleur. Mais cette technologie nécessite tout de même des outils d'analyse spécifiques. Par exemple, les intensités peuvent varier d'un individu à l'autre ou dépendre fortement des conditions extérieures [7]. D'autres effets, comme les changements de polarité thermique des objets ou l'inhomogénéité des corps humains, représentent des difficultés supplémentaires dans la segmentation des personnes. De plus, même si elles sont moins sensibles aux conditions d'illumination que les images couleurs, les émissions infrarouges restent dépendantes de certains facteurs comme la lumière du soleil sur les objets.
L'approche SKDA (Sequentiel Kernel Density Approximation) a déjà prouvé son efficacité pour la modélisation du fond [9] ou pour des algorithmes de suivi basés sur l'apparence [10]. Cette méthode tire sa robustesse de sa capacité à encoder plusieurs modes, et à s'adapter à des variations lentes au cours du temps, en intégrant de nouveaux échantillons et en délaissant les plus anciens. L'algorithme SKDA permet en outre d'obtenir une représentation compacte de l'information, puisque les modes proches les uns des autres peuvent être fusionnés grâce à une procédure de mean-shift, avec une complexité temporelle linéaire [10]. Dans le cas de notre capteur IR, il se peut qu'un léger déplacement des intensités du fond se produise, en raison par exemple de la présence de sources chaudes dans le champ de vision de la caméra. Cet effet indésirable peut conduire à de médiocres performances de l'algorithme de soustraction de fond. Pour surmonter cette difficulté, nous proposons une soustraction de fond en deux temps : après une première soustraction de fond, nous calculons l'offset entre la moyenne des modes du modèle SKDA, et la moyenne des pixels classés comme appartenant au fond, et nous appliquons cet offset lors d'une seconde soustraction de fond. Dans de nombreux cas difficiles, l'image après soustraction s'en trouve nettement améliorée.

3.2 Configuration du système d'acquisition

Le scénario qui nous intéresse ici se déroule sur le quai d'une gare où des personnes attendent un train. Pour le type de capteur utilisé, le choix sur les objectifs disponibles reste encore restreint et nous ne disposons que d'une focale de 25 mm. Par conséquent, pour couvrir une vaste étendue et prendre en compte les différentes contraintes du site, le capteur est placé à 10m en face du quai. Dans cette configuration, la profondeur du quai peut être considérée comme faible devant sa largeur. Nous l'approximons donc par un plan vertical dans l'espace 3-D, délimité verticalement par une ligne située à hauteur du sol (z=0), et une autre située à z=2m (voir FIG.3). Une seconde approximation nous permet de définir une correspondance directe entre les coordonnées dans ce plan 3-D et la région d'intérêt 2-D correspondante dans l'image: pour une position horizontale donnée dans l'image, la distance en pixels entre les deux lignes correspond à une hauteur réelle de 2m.

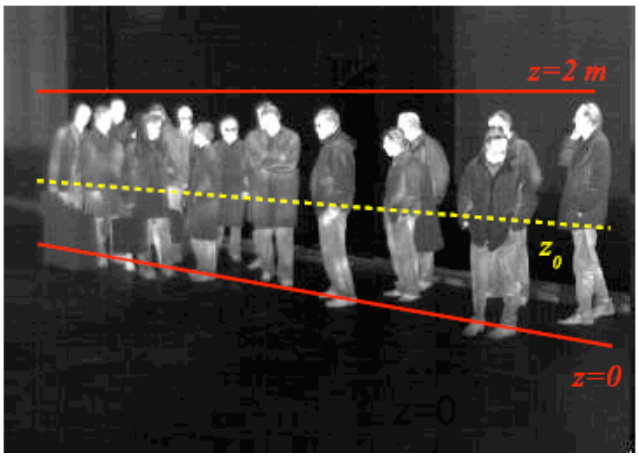

FIG. 3 : Le quai est modélisé par un plan 3-D du point de vue du capteur. La région d'intérêt est située entre les deux lignes z = 0 et z = 2m. La ligne en pointillés représente le seuil de hauteur z_0 utilisé pour détecter les personnes à terre.

3.3 Modélisation du corps humain

Pour représenter les personnes debout, nous utilisons un modèle géométrique 2-D. La tête est modélisée par une ellipse, et le torse et les jambes par deux rectangles verticaux (voir FIG.4). Si un tel modèle peut paraître simpliste, il présente l'avantage de nous éviter des processus complexes de projection et d'évaluations dans l'image. Des rectangles 2-D nous permettent par exemple d'utiliser des images intégrales [18]. Rappelons que notre objectif ici est d'avoir une bonne approximation de l'espace occupé par une personne debout, pour la distinguer des autres composantes présentes dans l'image (comme des personnes couchées au sol), et non de segmenter précisément toutes les parties du corps pour retrouver son attitude exacte, ce qui n'est pas l'objet de cet article.

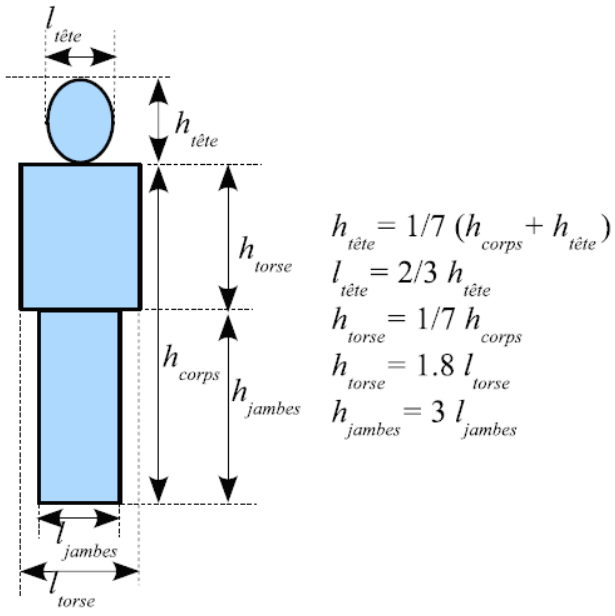

FIG. 4 : Modèle 2-D d'être humain avec une ellipse pour la tête et deux rectangles pour le torse et les jambes

3.4 Détection de la tête

A l'exemple de [20], des hypothèses sur les positions des têtes dans l'image sont générées par deux méthodes. La première est une détection des pics dans l'image obtenue après soustraction de fond. Un pic correspond à un maximum local vertical sur l'ensemble des blobs extraits. Nous définissons une région d'intérêt par une procédure de calibrage manuelle, qui nous donne une estimation de la taille d'une personne en fonction de sa position horizontale x dans l'image. Ce calibrage permet aussi de déterminer la taille de la zone de recherche des maxima locaux. La seconde méthode est basée sur une détection des têtes dans l'image infrarouge grâce à un modèle elliptique, comme décrit dans [3]. L'idée est que dans des images thermiques, les gradients les plus hauts sont observés autour des parties du corps les plus exposées comme le visage. Pour chaque position x du centre du modèle elliptiqueΓ, un score de concordance est calculé :

$$S_{\Gamma(x)} = \frac{1}{N_{x_i}} \sum_{x_i \in \Gamma(x)} \nabla I(x_i).n(x_i) \qquad (1)$$

où les x_i sont des points distribués sur l'ellipse, ∇I le gradient de l'image, et $n(x_i)$ la normale à $\Gamma(x)$ au point x_i. Les têtes sont recherchées uniquement à l'intérieur de la zone d'intérêt définie dans 3.2, et l'échelle du modèle elliptique est adaptée selon la position horizontale dans l'image.

Les hypothèses sur les positions des têtes sont en outre validées en calculant l'intersection entre le modèle 2-D de forme humaine correspondant à ces positions et la carte F obtenue par soustraction de fond. Pour accélérer le calcul de cette intersection, nous utilisons l'image intégrale de F pour les parties rectangulaires du modèle (le torse et les jambes). Toutes ces détections sont enfin fusionnées grâce à un algorithme de clustering séquentiel.

3.5 Détection de l'ensemble tête-épaules grâce à une cascade de classifieurs

Les gradients significatifs le long des contours de la silhouette des individus, et en particulier la forme de l'ensemble tête-épaules, représentent des informations pertinentes pour la détection de personnes. En complément de la méthode précédente, nous utilisons donc le résultat d'un détecteur "tête-épaules" basé sur des descripteurs locaux combinés dans une cascade de classifieurs [18]. En raison de la grande variabilité des apparences et des postures humaines, un descripteur robuste est nécessaire pour représenter les caractéristiques pertinentes. Nous avons utilisé des histogrammes de gradient (comme [5] et [2]) constitués de n cellules d'orientation et d'une cellule additionnelle représentant la quantité d'information contenue dans le support de l'histogramme. Après des normalisations en luminance, ces histogrammes sont calculés sur une grille dense (en position et en échelle), pour capturer de la manière la plus fine possible les caractéristiques de la forme "tête et épaules" que nous souhaitons reconnaître. Nos paramètres par défaut donnent des histogrammes de taille 900. L'utilisation d'une image intégrale permet de faciliter le calcul des valeurs du gradient et des votes. Nous avons observé que notre descripteur est plus performant avec 9 subdivisions sur l'orientation non-signée (tous les 20°entre 0°et 180°). Nous obtenons ainsi des vecteurs à 9000 composantes pour chaque forme représentée. L'utilisation d'orientations non-signées implique qu'il n'y a pas de distinction entre les différences zone claire/zone sombre et zone sombre/zone claire (ce qui est raisonnable si on considère la variabilité des apparences humaines : cheveux, peau, vêtements...). Pour réduire l'aliasing, nous effectuons un lissage des composantes de l'histogramme en attribuant une fraction x du vote à la cellule correspondante et une fraction 1-x à la cellule la plus proche, où $x \in [x_{\min}, 1]$. $x_{\min}$ dépend de l'angle seuil α_T au dessus duquel on considère qu'un vote se fait uniquement sur une cellule.

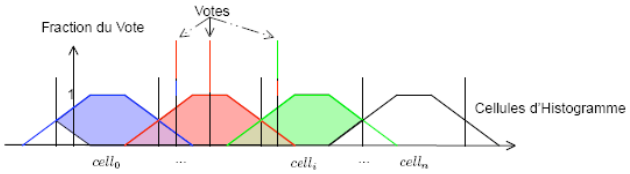

FIG. 5 : Histogramme des votes lissés

L'apprentissage est réalisé par une cascade de classifieurs [8, 18], où les classifieurs faibles sont de simples arbres de décision à un niveau (decision stump) sur les cellules de l'histogramme. L'objectif de cette approche et de réduire au maximum le nombre de zones d'intérêt candidates au fur à mesure de la cascade, de manière que la première couche de la cascade élimine la majorité des zones d'intérêt et que la dernière couche n'ait que quelques régions à évaluer. La figure 6 présente une vue d'ensemble de cette méthode. L'évolution de l'erreur et du taux de

détection au cours des différentes étapes de la cascade permet au détecteur d'obtenir au final de bons résultats. Une cascade à 10 étages permet en effet de parvenir à un taux de détection de 0.9 lorsque chaque étage possède un taux de 0.99 (puisque $0.99^{10} \approx 0.904$). De la même façon, un taux de faux positifs de 10^{-4} est quasiment atteint avec 10 étages ayant un taux de faux positifs de 40% ($0.04^{10} \approx 1.0 \times 10^{-4}$). Nous avons testé différents jeux de paramètres d'initialisation conduisant à des cascades de 9 à 12 étages. Les résultats de ces détecteurs diffèrent principalement sur le nombre de faux positifs, les taux de détection étant en revanche assez similaires.

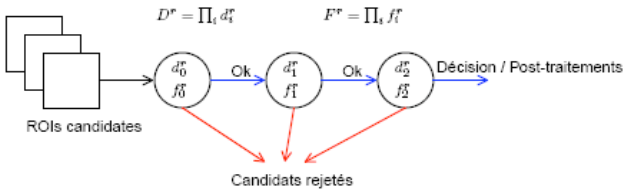

FIG. 6 : Cascade de classificateurs, où D^r est le taux de détection et F^r le taux de faux positifs.

Lors de la procédure d'apprentissage, nous avons utilisé $n_+ = 3000$ exemples positifs et $n_- = 10000$ exemples négatifs *difficiles*. Les exemples difficiles sont obtenus sur une séquence à l'aide d'un classificateur simple (quelques étages), entrainé sur des exemples négatifs choisis aléatoirement. La procédure de boosting sélectionne les composantes pertinentes (correspondant à des classifieurs faibles) pour construire un classificateur fort, lui-même utilisé pour choisir n_- exemples négatifs pour le prochain étage de la cascade. La dernière couche entrainée est évaluée sur un ensemble de validation, et si le résultat n'est pas satisfaisant, une nouvelle couche est ajoutée à la cascade. Pour améliorer la procédure et la rendre plus robuste, une fois que les classifieurs faibles ont été sélectionnés pour l'étage courant de l'algorithme, une autre boucle permet d'ajuster les poids des différents classifieurs.

3.6 Raffinement de la segmentation par un échantillonnage MCMC

Une fois que les têtes ont été localisées précisément par notre détecteur, nous procédons à une étape de segmentation dont le but est d'ajuster la position et la forme du modèle de corps humain pour les personnes debout. Cette procédure permet par la suite une meilleure analyse des blobs restants. Le problème de segmentation peut être formulé sous la forme d'une estimation d'un maximum a posteriori (MAP) :

$$\Theta^* = \arg\max_{\Theta} p(\Theta/F) \qquad (2)$$

où $\Theta = \{\theta_i\}$ est l'ensemble des paramètres des différents modèles humains, et F est la carte obtenue après soustraction du fond. D'après la règle de Bayes, la probabilité a posteriori peut être décomposée en un terme de vraisemblance et une probabilité a priori :

$$p(\Theta/F) \propto p(F/\Theta)\, p(\Theta) \qquad (3)$$

Les paramètres de chaque individu i sont $\theta_i = \{\Delta x_i, h_i, f_i\}$, où Δx est la translation horizontale du corps par rapport à sa position initiale, h sa hauteur et f sa corpulence (rapport de la largeur sur la hauteur).
Comme les paramètres des modèles des différents individus sont indépendants les uns des autres, on peut supposer que la probabilité a priori jointe est le produit des probabilités a priori pour chaque individu :

$$p(\Theta) = \prod_{i=1}^{N} p(\theta_i) \qquad (4)$$

où N est le nombre de personnes debout détectées.
Pour un individu i, probabilité a priori est :

$$p(\theta_i) = p(\Delta x_i)\, p(h_i)\, p(f_i) \qquad (5)$$

où $p(\Delta x_i)$ est une distribution gaussienne $N(0, \sigma_{\Delta x})$ tronquée sur l'intervalle [-0.4, 0.4], $p(h_i)$ est une distribution uniforme sur l'intervalle $[h_{i0} - 0.3, h_{i0} + 0.3]$(où h_{i0} est la taille initiale du modèle), et $p(f_i)$ est une distribution uniforme sur l'intervalle [0.9, 2.2]. Comme plusieurs personnes peuvent s'occulter les unes les autres, la vraisemblance jointe ne peut pas s'exprimer comme le produit des vraisemblances de chaque hypothèse d'individu. Nous utilisons donc une vraisemblance basée sur le nombre de pixels "mal classés", c'est-à-dire N_{01}, le nombre de pixels qui appartiennent à la carte F obtenue après soustraction du fond mais qui ne sont dans aucun modèle humain, et N_{10}, le nombre de pixels qui sont dans un modèle de personne mais qui ne sont pas dans F:

$$p(F/\Theta) = \sigma(\lambda_{01} \frac{\Delta N_{01}}{N}).\sigma(\lambda_{10} \frac{\Delta N_{10}}{N}) \qquad (6)$$

où $\sigma(x) = 1/(1+e^{-x})$ est la fonction sigmoïde, ΔN_{01} (resp. ΔN_{10}) est la différence entre la valeur courante et la valeur initiale de N_{01} (resp. N_{10}), et λ_{01} et λ_{10} sont deux coefficients de pondération dépendant de la taille des êtres humains dans l'image.
Pour maximiser une telle fonction, les méthodes d'échantillonnage nous fournissent un moyen simple d'explorer l'espace des états possibles et d'évaluer la solution optimale avec une grande robustesse vis-à-vis des maxima locaux. Les paramètres optimaux Θ^* sont calculés par une approche Monte Carlo par chaînes de Markov (MCMC) à l'exemple de [20] pour la segmentation de personnes, et de [11] dans le contexte du suivi multi-cibles. L'algorithme de Metropolis-Hastings est une technique efficace pour échantillonner une distribution quelconque, en construisant séquentiellement une chaine de Markov qui

converge vers cette distribution. Nous avons utilisé comme distribution instrumentale une distribution gaussienne. Les principales étapes de l'algorithme sont les suivantes :

– Initialiser les modèles de personnes d'après les résultats de la détection de tête, avec les paramètres définis en 3.3 et une échelle définie par la position horizontale x dans l'image

– Pour chaque échantillon :

1. choisir un individu i au hasard,
2. à partir de l'état courant θ_i^t, prédire un nouvel état θ_i^{t+1} avec la distribution instrumentale
3. estimer la nouvelle probabilité a posteriori $p^{t+1}(\Theta/F)$
4. calculer le taux d'acceptation $r = \dfrac{p^{t+1}(\Theta/F)}{p^t(\Theta/F)}$
5. si r>1 le nouvel état θ^{t+1} est accepté avec une probabilité r

Une fois que les échantillons de la distribution postérieure sont générés, une estimation de l'état est obtenue en calculant la moyenne pondérée des paramètres des différents échantillons.

4. Détection d'une menace

Pour chaque image de la vidéo, notre système de détection des menaces peut produire quatre sorties possibles : scène *vide* lorsqu'aucune personne n'a été détectée, *normal* si des individus debout ont été détectés et personne n'est à terre, *avertissement* ou *alarme*, selon le niveau de confiance, dans le cas où des personnes allongées par terre ont été détectées. La décision est prise en calculant une probabilité de menace associée à l'événement *personne à terre*, $p_t(LD)$ et en la comparant à deux seuils, un seuil d'avertissement τ_w et un seuil d'alarme τ_A tels que $0 < \tau_w < \tau_A < 1$.

4.1 Détection des personnes allongées par terre

Pour déterminer si des personnes sont à terre, notre algorithme se base sur :

– la hauteur des têtes détectées : en dessous de $z_0 = 1m$, la personne est classée comme couchée (voir 6 pour les personnes de petite taille),

–l'analyse des blobs restants, une fois les personnes debout supprimées. Les personnes debout qui ont été segmentées sont donc supprimées de la carte obtenue après soustraction de fond. Des opérations de morphologie mathématique sont ensuite appliquées à cette image binaire pour éliminer les régions les plus fines. Dans les images infrarouges, les corps humains ont généralement une texture plus hétérogène que les objets inertes, comme les valises par exemple. Les blobs restants sont finalement analysés et classés comme des personnes allongées ou d'autres objets, en utilisant un critère sur la distance au sol et la texture, caractérisée par la variance locale calculée sur un voisinage de 3x3 pixels.

4.2 Probabilité d'une menace

On note N_l le nombre d'hypothèses sur les personnes à terre à un instant t donné. La probabilité de menace associée à l'événement *personne à terre*, $p_t(LD)$ peut s'exprimer comme le produit de trois probabilités. Le premier terme $p_t^{N_l}(LD)$ est lié au nombre d'hypothèses sur les personnes à terre détectées : plus le nombre d'hypothèses est grand, plus la probabilité de menace est élevée. La deuxième probabilité $p_t^{X_l}(LD)$ dépend des positions estimées de ces hypothèses par rapport au sol : le niveau de confiance augmente lorsque la distance moyenne au sol diminue. Le troisième terme $p_t^{fl}(LD)$ exprime la fréquence de détection dans une fenêtre temporelle. Un historique de la détection d'événements sur une fenêtre temporelle de largeur h_w est conservé, et on compte le nombre d'occurrences n_l de l'événement *personne à terre* détectées. La probabilité résultante peut s'écrire :

$$p_t(LD) = \underbrace{\frac{1}{1+e^{-\lambda_1 N_l}}}_{p_t^{N_l}(LD)} \cdot \underbrace{\frac{1}{1+e^{-\frac{\lambda_2}{N_l}\sum_{i=1}^{N_l}\frac{z_i}{z_0}}}}_{p_t^{X_l}(LD)} \cdot \underbrace{\frac{n_l}{h_w}}_{p_t^{fl}(LD)} \qquad (7)$$

où λ_1 et λ_2 sont deux coefficients de pondération, fixés empiriquement à λ_1=1 et λ_2=2, et où z_i représente la distance des hypothèses au seuil de hauteur z_0.

5. Résultats expérimentaux et discussion

Le détecteur de menace a été testé à de nombreuses reprises dans le cadre du projet. Nous présentons ici les résultats obtenus sur deux séquences longues représentatives (3351 et 7342 images respectivement). La dimension des images est de 384x272 pixels. Dans ces séquences, un groupe de personnes pénètrent dans la zone vide, et restent debout sur le quai. A un instant donné, certaines d'entre elles tombent par terre, d'autres restent debout. Les résultats obtenus lors des différentes étapes de l'algorithme sont illustrés dans les figures qui suivent. La figure 7 présente les résultats de soustraction de fond obtenus avec la méthode SKDA.

FIG. 7 : Résultats de la soustraction de fond.

Sur toutes les images traitées, aucun des individus présents n'a été supprimé lors de la soustraction de fond. En revanche, quelques fausses détections ont été observées, mais celles-ci ont pu être filtrées lors des étapes suivantes. La figure 8 présente les résultats de la détection de têtes. Le détecteur s'est révélé très robuste étant donné la complexité des scènes à traiter en termes de densité humaine, d'occultations et de variabilité des postures.

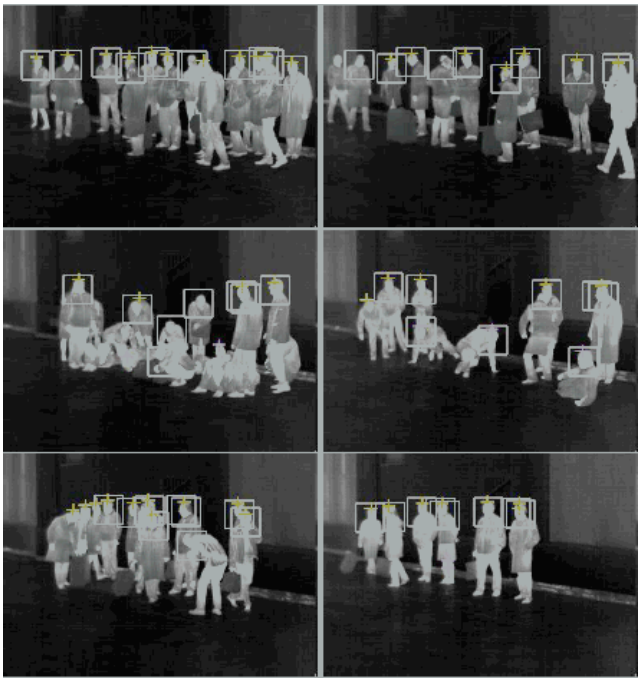

FIG. 8 : Résultats de la détection de têtes. Les croix indiquent les têtes détectées par la détection des pics et le détecteur de formes elliptiques (en jaune au dessus du seuil sur la hauteur, et en magenta en dessous), les boîtes représentent les résultats de détection de la cascade de classificateurs. La dernière image montre un exemple de fausse détection sur les jambes, qui présentent un fort gradient.

Le détecteur de formes elliptiques et la cascade de classifieurs parviennent à détecter des têtes même lorsqu'une autre personne placée en arrière plan vient altérer les contours de la tête ou lorsque la personne s'est penchée. On peut aussi souligner la complémentarité des différents détecteurs dans les cas difficiles. Des fausses détections en dessous du seuil sur la hauteur ont été observées dans seulement 14 des 10639 images. Le nombre d'hypothèses sur les têtes situées au-dessus du seuil est une estimation du nombre de personnes debout dans la scène, la précision de l'estimation dépendant naturellement de la densité et du niveau de chevauchement. Nous avons représenté sur un graphique le nombre d'hypothèses générées pour les têtes sur une séquence de 2555 vues. Au début de cette séquence, les 15 personnes présentes sont debout. Après environ 1150 vues, 9 personnes s'en vont et durant les 600 autres vues, 6 individus restent debout sur le quai (voir FIG.9).

FIG. 9 : Estimation du nombre de personnes debout. Au début de la séquence, 15 personnes sont debout. 9 quittent la zone d'intérêt, et 6 restent.

Les trois principales phases de la séquence sont visibles sur le graphique. Comme on pouvait s'y attendre, l'erreur d'estimation augmente avec la densité des personnes présentes, mais les résultats restent cohérents avec la vérité terrain. La figure 10 présente le résultat de la segmentation de personnes debout avec l'approche bayésienne et le modèle 2-D du corps humain. Les paramètres optimaux obtenus après l'échantillonnage MCMC permettent au modèle de mieux s'ajuster à l'apparence des individus. En particulier, la hauteur et la corpulence de la personne restent correctement estimées et l'inclinaison du corps d'une personne peut être compensée dans l'image par une translation du corps par rapport à la tête.

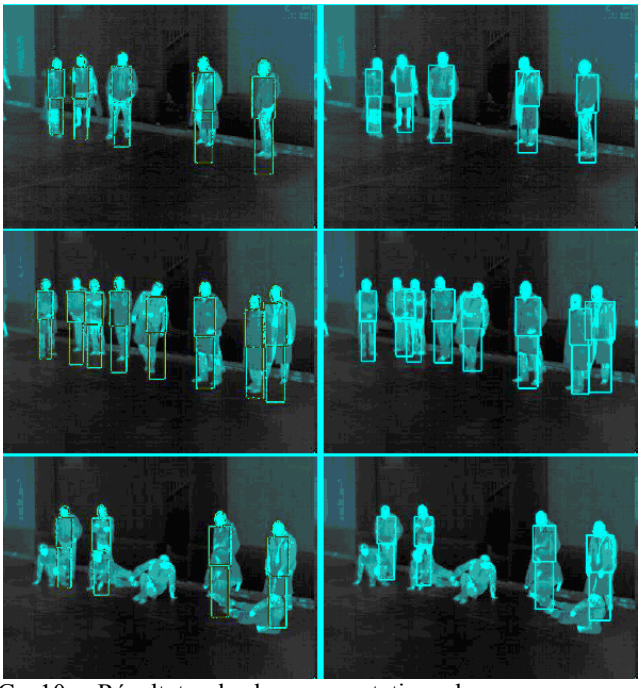

FIG. 10 : Résultats de la segmentation des personnes debout. Colonne de gauche : position initiale, colonne de droite : segmentation finale après échantillonnage MCMC

Une fois que les corps des personnes debout ont été supprimés, les blobs restants et proches du sol sont classés comme étant des personnes debout ou d'autres objets, comme indiqué sur la figure 11. Les cas de non détection sont dus à la forte densité locale et à un chevauchement extrême.

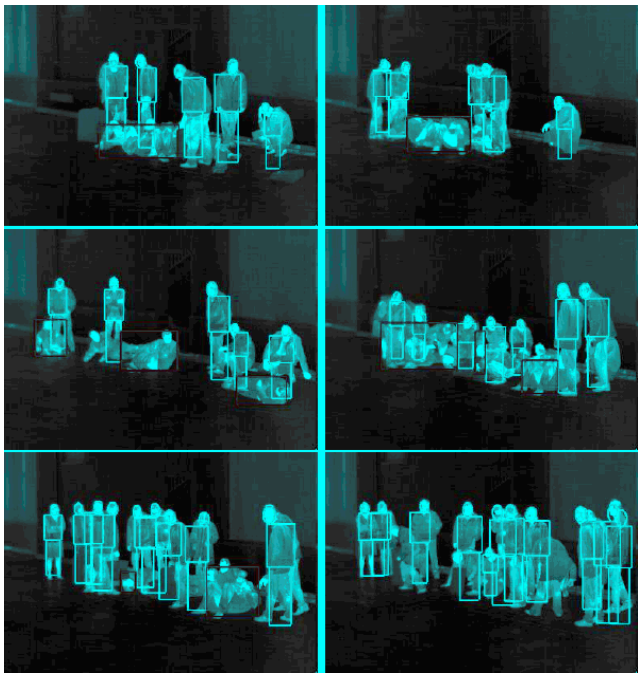

FIG. 11 : Détections des chutes. La dernière image illustre un cas de non détection.

La table 1 donne les résultats de la détection des menaces : les différentes vues sont classées comme vide, normale ou avertissement/alarme. L'algorithme donne des résultats satisfaisants puisque l'estimation est cohérente avec les nombres de vues de la vérité terrain.

D'un point de vue temporel, si nous représentons les niveaux d'alarme au cours du temps (FIG. 12), on observe une bonne concordance entre la sortie de l'algorithme et la vérité terrain sur les séquences prétraitées. On peut noter un léger décalage de quelques images entre le début de la menace dans la vérité terrain et l'activation de l'alarme par notre système. Ceci est dû à la largeur de la fenêtre temporelle utilisée pour calculer la probabilité de fréquence. Pour cette application, un retard de quelques secondes pour le déclenchement de l'alarme est tout à fait acceptable surtout s'il permet de réduire les risques de fausses alarmes. Il reste néanmoins certains aspects à améliorer. Au début de la séquence 2, de fausses alarmes apparaissent brièvement. De plus, l'alarme n'est pas activée de façon continue durant la période critique à cause des cas intermittents où les personnes à terre ne sont pas détectées. Le lissage temporel de la détection de menace pourrait être amélioré grâce à un filtrage à plus long terme.

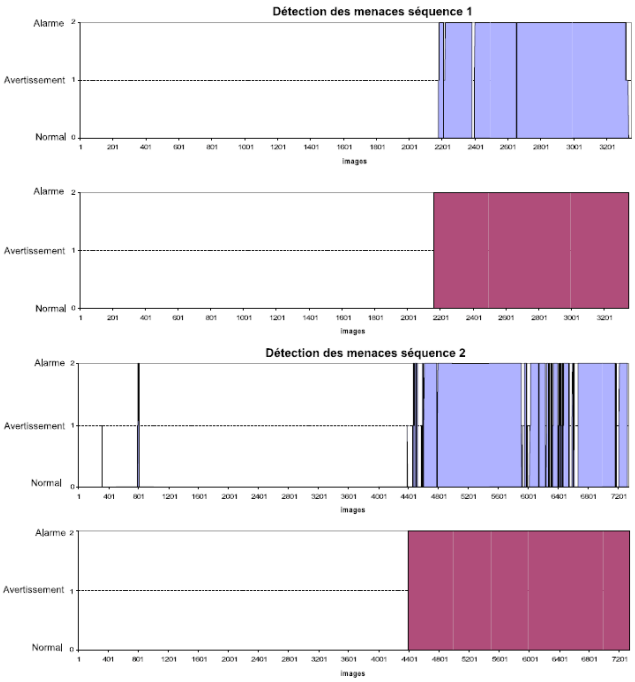

FIG. 12 : Résultats de la détection de menaces.
1ère ligne : sortie de l'algorithme pour la séquence 1,
2ème ligne : vérité terrain pour la séquence 1,
3ème ligne : sortie de l'algorithme pour la séquence 2,
4ème ligne : vérité terrain pour la séquence 2.

En termes de rapidité, avec un code C++ pouvant être encore sérieusement optimisé, notre algorithme traite approximativement 2-3 images par secondes sur un PC conventionnel Pentium IV 3Ghz, 1.5Gb RAM. Cette fréquence de traitement est largement suffisante pour l'application visée.

TAB. 1 : Résultats de la détection de menaces (en nombre de vues).

Seq 1	*Vide*	*Normal*	*Avertissements/Alarmes*
Estimation	0	2214	1137
Vérité terrain	0	2161	1190
Seq 2	*Vide*	*Normal*	*Avertissements/Alarmes*
Estimation	0	4784	2557
Vérité terrain	0	4394	2948

6. Conclusion et perspectives

Dans cet article, nous avons démontré les capacités de notre système à analyser des scénarios complexes de détection de menaces dans des images infrarouges, comme la détection des personnes tombées à terre dans une foule. Les résultats expérimentaux montrent la robustesse de la méthode, puisque le taux de fausses alarmes est bas, et peu d'alarmes attendues ont été manquées. Les performances de la détection pourraient être améliorées en intégrant un lissage temporel, à la fois dans le processus de segmentation (suivi visuel) et dans la sortie du module de détection de menace (long terme). Une autre amélioration possible serait l'enrichissement du modèle de silhouette pour augmenter la précision de la segmentation à un faible coût calculatoire. Un meilleur modèle permettrait de distinguer les personnes de petites tailles des personnes agenouillées. Un modèle multicouche des groupes de personnes pourrait également permettre de mieux gérer les occultations. De plus, les résultats obtenus dans cette étude dépendent largement de la position du capteur infrarouge et de son champ de vision. Idéalement, une vue de dessus et un champ de vision plus large diminueraient le chevauchement entre les individus.

Références

[1] E.L.Andrade, S.Blunsden and R.B. Fisher. *Hidden markov models for optical flow analysis in crowds*. In Proc. IEEE Int. Conf. On Pattern Recognition, Vol.1, pages 460–463, Los Alamitos, CA, USA, 2006.

[2] J.Begard, N.Allezard, and P.Sayd. *Real-time humans detection in urban scenes*. In BMVC, Warwick, UK, 2007.

[3] S.Birchfield. *Elliptical head tracking using intensity gradients and color histograms*. In Proc. IEEE Conf. on Computer Vision and Pattern Recognition, pages 232–237, 1998.

[4] C.O.Conaire, E.Cooke, N.O'Connor, N.Murphy, and A. Smearson. *Background modelling in infrared and visible spectrum video for people tracking*. In Proc. IEEE Conf. on Computer Vision and Pattern Recognition Workshop, page 20, Washington, DC, USA, 2005.

[5] N.Dalal and B.Triggs. *Histograms of oriented gradients for human detection*. In *CVPR*, volume II, pages 886–893, 2005.

[6] A.C.Davies, J.H.Yin, and S.A.Velastin. *Crowd monitoring using image processing*. Electronics and Communications Engineering Journal, 7(1) :37–47, 1995.

[7] J.W. Davis and V.Sharma. *Robust background-subtraction for person detection in thermal imagery*. In Proc. IEEE Conf. on Computer Vision and Pattern Recognition Workshop, volume 8, page 128, Washington, DC, USA, 2004.

[8] J.Friedman, T.Hastie, and R.Tibshirani. *Additive logistic regression : a statistical view of boosting.* Technical report, Dept. of Statistics, Stanford University, August 1998.

[9] B.Han, D.Comaniciu, and L.Davis. *Sequential kernel density approximation through mode propagation : applications to background modeling*. In Proc. of the 2004 Asian Conference on Computer Vision, 2004.

[10] B.Han and L.Davis. *On-line densitybased appearance modeling for object tracking*. In Proc. IEEE Int. Conf. on Computer Vision, pages 1492–1499, Washington, DC, USA, 2005.

[11] Z.Khan, T.Balch, and F.Dellaert. *MCMC-based particle filtering for tracking a variable number of interacting targets.* IEEE Trans. on Pattern Analysis and Machine Intelligence, 27(11) :1805–1918, 2005.

[12] P. Kilambi, O.Masoud, and N.Papanikolopoulos. *Crowd analysis at mass transit sites*. In ITSC '06, Intelligent Transportation Systems Conference, pages 753 – 758, Toronto, Canada, September 2006.

[13] A.Marana, S.Velastin, L.Costa, and R.Lotufo. *Automatic estimation of crowd occupancy using texture and nn classification*. Safety Science, 28(3): pp.165–175, 1998.

[14] L.Davis and H.Nanda. *Probabilistic template based pedestrian detection in infrared videos*. In IEEE Intelligent Vehicle Symposium, pages 18–20, June 2002.

[15] A.J.Policastro and S.PGordon. *The use of technology in preparing subway systems for chemical/ biological terrorism*. In Commuter Rail/Rapid Transit Conference Proceedings, 1999.

[16] P. Reisman, O. Mano, S. Avidan, and A. Shashua. *Crowd detection in video sequences*. In IEEE Intelligent Vehicles Symposium, June 2004.

[17] Gandhi T. and Trivedi M.M. *Pedestrian collision avoidance systems: A survey of computer vision based recent studies*. In Intelligent Transportation Systems Conference, pages 976–981, Toronto, Canada, September 2006.

[18] P.Viola and M.Jones. *Robust real-time object detection.* In International Workshop on Statistical and Computational Theories of Vision Modeling, Learning, Computing and Sampling, July 2001.

[19] B. Wu and R. Neviata. *Detection of multiple, partially occluded humans in a single image by Bayesian combination of edgelet part detectors.* In ICCV, 2005.

[20] T.Zhao and R.Nevatia. *Bayesian human segmentation in crowded situations.* In Proc. IEEE Conf. on Computer Vision and Pattern Recognition, pages 459–466, June 2003.