Vocabulaire Visuel Hiérarchique pour la détection et le suivi de piétons en utilisant l'infrarouge lointain

Bassem BESBES, Alexandrina ROGOZAN, Abdelaziz BENSRHAIR

Laboratoire d'Informatique, du Traitement de l'Information et des Systèmes
INSA de Rouen Avenue, de l'université, BP 8, 76801 Saint-Étienne-du-Rouvray Cedex, France
bassem.besbes@insa-rouen.fr, alexandrina.rogozan@insa-rouen.fr
abdelaziz.bensrhair@insa-rouen.fr

Résumé – La mise en place d'un système de détection de piétons embarqué à bord d'un véhicule est confrontée à des contraintes temps réel et différents problèmes liés principalement à la variabilité de l'apparence et de la forme des piétons. Dans cet article, nous nous intéressons à la mise en place d'un système bien adapté aux problématiques de détection de piétons dans les images infrarouge-lointain. Ce système est basé sur la mise en correspondance entre des points d'intérêts avec des motifs regroupés dans un Vocabulaire Visuel Hiérarchique. La structure hiérarchique permet à la fois d'accélérer le temps d'appariement et d'améliorer les performances de détection. Les résultats de détection de piétons obtenus sont très satisfaisants et mettent en évidence le bon comportement du système surtout face aux problèmes de changements d'échelle et d'occultations partielles.

Abstract – The implementation of an embedded vision-based pedestrian detector is confronted with real-time constraints and various problems related to the large variability in the shape and appearance of pedestrians. In this paper, we present a novel pedestrian detector that exploits the specific characteristics of Far-Infrared images. The system is based on the matching of interest points extracted from images with patterns stored in a hierarchical codebook. The hierarchical structure allows in the one hand to accelerate the matching processing time and in the other hand to improve the performance of detection. The experimental evaluation shows that our detector is robust since it allows to deal with scale changes and partial occlusion problems.

1 Introduction

La détection et le suivi de piétons ont fait l'objet de nombreux travaux pour une multitude d'applications. Dans le cadre de l'aide à la conduite automobile, ce sujet est assez difficile et assez important car le piéton est l'objet le plus vulnérable présent dans une scène routière. La complexité de la tâche provient d'une part, de la grande variabilité de formes et d'apparences des piétons ; et de l'autre part, des changements d'échelle, d'orientation et des variations des points de vue. Des solutions basées sur l'appariement de points d'intérêt (POI) sont très employées actuellement et peuvent contribuer à résoudre une partie de ces problèmes [3, 4, 5]. L'idée consiste à extraire des points d'intérêt robustes et des descripteurs locaux autour, permettant des appariements résistants à la rotation, la translation et la mise à l'échelle. Les solutions basées sur les descripteurs locaux sont particulièrement appropriées aux images infrarouges (IR) où la variance des apparences des piétons, bien contrastés par rapport au fond, est plutôt limitée par rapport aux images visibles [8, 9]. Ces solutions permettent de s'affranchir des limites des méthodes de détection classiques employées en IR comme la soustraction de fond, le seuillage et l'extraction de cartes de symétrie [2, 6].
De ce fait, nous avons procédé tout d'abord à identifier des points d'intérêt SURF [5] dans des zones à fort contraste. Le choix de SURF est justifié, d'une part, car c'est un rapide détecteur et robuste descripteur. D'autre part, car il est bien adapté aux caractéristiques de l'IR puisqu'il permet d'extraire des régions claires dans l'image (ayant une valeur négative de Laplacien). En revanche, les images IR sont moins nettes et les régions correspondantes aux piétons ont des intensités non uniformes. De plus, les images de scènes routières urbaines prises en utilisant l'IR lointain présentent de fréquentes occultations qui compliquent le processus de détection. Afin de surmonter ces difficultés, nous proposons, dans un premier temps, de détecter les régions bien contrastées correspondantes aux têtes de piétons. Ces régions sont identifiées en utilisant un Vocabulaire Visuel Hiérarchique, établi en apprentissage, regroupant les caractéristiques locales de ces régions. Ensuite, le système de détection proposé procède par la construction de fenêtres d'intérêt qui sont par la suite validées par un SVM linéaire [9]. Finalement, les piétons détectés feront l'objet d'un processus de suivi basé sur l'appariement temporelle des descripteurs SURF.

2 Vocabulaire Visuel Hiérarchique

La construction d'un Vocabulaire Visuel (VV), d'une façon générale, consiste à établir des catégories de motifs qui ont une particularité commune. Dans ce contexte, la construction d'un

VV revient à quantifier l'espace des descripteurs des POI. Il est à mentionner que les descripteurs SURF sont basés sur des sommes de réponses d'ondelettes de Haar et sont représentés sous la forme d'un vecteur caractéristique de taille fixe. Afin de créer le VV, nous avons rassemblé et regroupé les descripteurs SURF localisés dans des régions de tête à partir d'un ensemble de fenêtres englobantes faisant partie des données d'apprentissage. Pour chaque POI retenu, nous proposons d'associer un paramètre (r) représentant le ratio entre l'échelle (ρ) à laquelle le POI a été extrait, et la distance (d) à la plus proche bordure de la fenêtre qui englobe le piéton (figure 1). L'enregistrement de ce paramètre permettra, dans la phase de détection, de générer une fenêtre autour d'un POI ayant un descripteur similaire (ayant des caractéristiques locales proches).

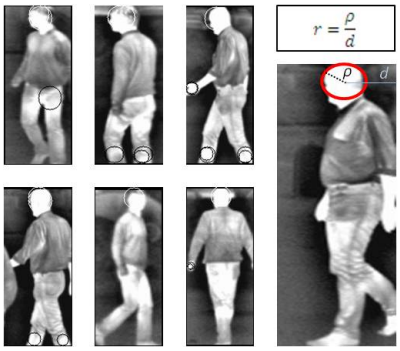

FIGURE 1 – Extraction de POI SURF localisés dans des régions de têtes (cercles en blanc) et l'enregistrement du rapport entre l'échelle et la distance à la plus proche bordure

Le VVH est représenté comme un arbre n-aire dont chaque élément représente un cluster caractérisé par : (a) un vecteur descripteur moyen (centroïde), (b) une valeur moyenne des paramètres r associés à chaque point contenu dans le cluster et (c) un rayon dont la valeur présente la distance Euclidienne entre le centroïde et le descripteur les plus éloigné dans le cluster. Chaque niveau hiérarchique résulte de l'application de l'algorithme de clustering agglomératif (RNN [7]) avec un seuil spécifique. Le seuil initial de clustering est choisi de manière à minimiser une fonction d'évaluation (F) qui implique le taux de clusters unitaires et le taux maximal de points contenus dans un cluster (Eq1). Le seuil maximum correspond, quant à lui, à la valeur minimale qui permet de regrouper tous les descripteurs dans un seul nœud (racine).

Soit N désigne le nombre total de clusters dans le VVH. N inclut N_u clusters unitaires ($n_i = 1$) and N_{nu} non unitaires.

$$F(t_1, l) = (Q_1) \times (Q_2) = (\frac{N_u}{N}) \times (\frac{\max_{j \in 1..N}(n_j)}{\sum_{k=1}^{N} n_k}) \quad (1)$$

Q_1 tend à maximiser le nombre de clusters non unitaires.
Q_2 assure un certain équilibre entre les tailles des clusters.

Pour les niveaux intermédiaires, nous avons choisi d'augmenter le seuil du bas vers le haut en doublant le pas entre deux niveaux consécutifs. Il est très important de mentionner que la profondeur de l'arbre est choisie également de manière à minimiser la fonction F. Ainsi, l'algorithme proposé ne nécessite pas la connaissance préalable de seuils et permet la construction d'un VVH optimal d'une façon automatique.

3 Détection et suivi de piétons

Le système de détection proposé se décompose en trois étapes principales : détection de têtes, construction et validation de fenêtres d'analyse (FI) et finalement un suivi temporel sur les piétons détectés.

3.1 Détection de têtes

La détection des têtes se base sur la mise en correspondance des POI SURF extraits d'une image de test avec le VVH construit en apprentissage. Ce processus est accéléré par l'exploitation de la représentation hiérarchique. En effet, une exploration partielle de l'arbre est généralement suffisante : lors de la mise en correspondance, il n'est pas nécessaire d'examiner les sous-arbres dont le nœud père n'a pas été activé. Un nœud s'active si la distance Euclidienne entre le descripteur du POI et le centroïde du cluster désigné est inférieure à son rayon. Par conséquent, la mise en correspondance est mise en œuvre en appliquant un simple algorithme itératif de parcours en profondeur. Un score d'appariement ($S_{i,k}$) est associé pour le plus profond nœud activé (i) dans chaque sous branche de l'arbre (Eq2). Ce score dépend à la fois de la distance entre les descripteurs (du POI f_k et du centroide C_i) et de la valeur du rayon du cluster.

$$S_{i,k} = \frac{R_i - d(f_k, C_i)}{R_i} \quad (2)$$

Après avoir sélectionné le cluster maximisant S, une fenêtre d'intérêt est déterminée en fonction du paramètre r associé au cluster et l'échelle du POI. Enfin, les fenêtres se chevauchant sont regroupées en adaptant le même algorithme de clustering agglomératif (RNN) utilisé pour créer le VV. Les extensions ajoutées consiste, tout d'abord, à considérer le plus proche voisin ($F2$) d'une fenêtre ($F1$) celui qui maximise le taux de chevauchement (rapport intersection/union). En deuxième temps, la fusion de ($F1$) et de ($F2$) produit une nouvelle fenêtre (voir figure 2) qui prend une position moyenne pondérée par le score de ($F1$) et de ($F2$).

FIGURE 2 – Regroupement des fenêtres se chevauchant

3.2 Construction et analyse des fenêtres d'intérêt

Les fenêtres d'analyse qui englobent les piétons sont estimées, en premier lieu, d'une façon grossière en utilisant le rapport hauteur/largeur. Ensuite, les positions des bas des fenêtres sont affinées par la recherche de la ligne qui contient plus de contours horizontaux. Dans [9], nous avons présenté une méthode rapide et fiable qui combine entre un SVM linéaire avec un VVH. Dans ce travail, nous avons utilisé le même classifieur qui opère sur le même VVH afin d'éliminer les fausses détections. En d'autres termes, les résultats de mise en correspondance entre les POI et le VVH établis dans l'étape de détection ont été réutilisés pour constituer l'entrée du classifieur SVM. Dans la figure 3 nous illustrons les résultats de détection de piétons obtenus après l'exécution de chaque étape de l'algorithme proposé.

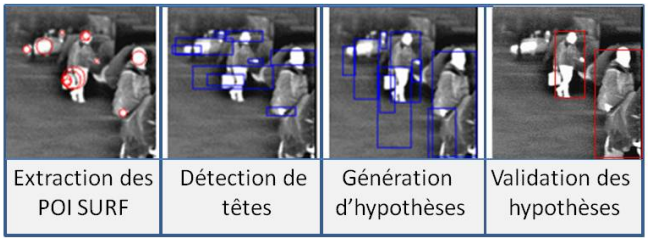

FIGURE 3 – Illustration des résultats de détection obtenus après chaque étape de l'algorithme proposé

3.3 Suivi de piétons

Un algorithme simple de suivi, basé sur la mise en correspondance temporelle des descripteurs SURF, est enclenché pour les piétons détectés. En effet, nous effectuons un appariement des points SURF entre deux images successives. Chaque couple de points appariés vote pour la nouvelle position de l'obstacle avec un score calculé en fonction de la distance entre les descripteurs. Ces votes ensuite sont interprétés par l'algorithme de Mean Shift [10, 1] mais en 3D (2D pour la position du centre et 1D pour la valeur de l'échelle). L'avantage de cet algorithme est qu'il est très rapide et très peu coûteux. Il permet de trouver les coordonnées de la fenêtre englobante (position et échelle) dans une nouvelle image de manière itérative jusqu'à ce qu'il y a convergence ou jusqu'à ce qu'un nombre maximum d'itérations est atteint. La figure 4 illustre le principe de l'algorithme de suivi proposé.

4 Résultats expérimentaux

Les expérimentations ont été menées sur des images provenant de deux séquences d'IR lointain qui ont été extraites d'un système appelé Tetravision [6]. Les caméras infrarouges utilisées sont sensibles aux longueurs d'ondes se situant entre 7 et $14\mu m$. La figure 5 présente des exemples de détection en présence de quelques occultations partielles. Il est clair que la majorité des piétons sont correctement détectés, indépendamment de leurs résolutions.

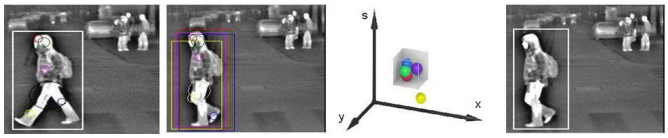

FIGURE 4 – Illustration du principe de l'algorithme de suivi. La première image contient un piéton détecté et un ensemble de POI entourés par des cercles, dont les rayons correspondent à leurs valeurs d'échelles. Dans l'image qui suit, chaque région d'intérêt est construite après l'appariement temporel des descripteurs. Chaque couple de POI apparié vote pour la position et l'échelle du piéton dans l'image suivante. L'ensemble des votes sont traités en 3D par l'algorithme de Mean Shift qui fournit en sortie les coordonnées optimales de l'emplacement du piéton (dernière image)

(a)

(b)

FIGURE 5 – Exemples de détections dans deux séquences d'IR lointoin. Toutes les images ont été traitées à leur résolution d'origine (320×240 pixels). Les résultats confirment la précision du système de détection même en présence d'occultations partielles.

Le tableau 1 récapitule les résultats obtenus et met en valeur la grande variabilité des résolutions des images des piétons. Le système de détection et de suivi a été évalué en s'appuyant aux mesures des taux de précision et de rappel. Le tableau montre que l'algorithme proposé retourne des résultats satisfaisants avec un bon compromis de précision-temps de calcul.

TABLE 1 – Résultats de détection en IR lointain

Séquence	nombre de piétons	plage des résolutions	résolution moyenne	écart-type des résolutions	Taux d'occultation	Précision %	Rappel %	Temps de calcul (ms)
N.1 (Fig5.a)	454	[320,32640]	7926.12	5019.54	17.4	0.86	0.88	102
N.2 (Fig5.b)	366	[133,24639]	6614.51	5525.17	13.66	0.75	0.74	104

Les performances de l'algorithme de détection et de suivi sont largement dépendants de la profondeur du VV. Dans la section 2, nous avons proposé une fonction (F) qui permet d'évaluer la pertinence de la structure du VV. Durant la phase de validation, Il a été constaté qu'un VV de profondeur 5 permet de maximiser la valeur de F. Nous présentons dans la figure 6 l'évolution des scores de détection en fonction de la profondeur du VVH.

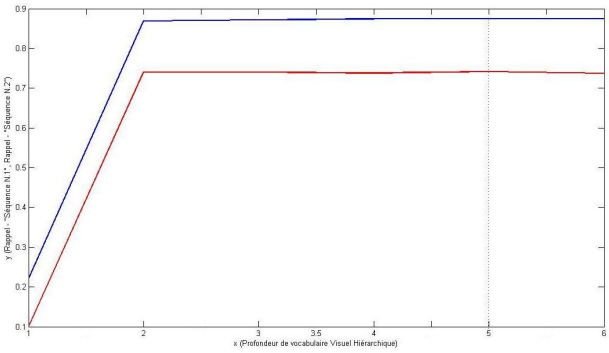

FIGURE 6 – L'évolution des valeurs de Rappel Vs profondeur du Vocabulaire Visuel Hiérarchique

Dans la figure 6, on observe que les meilleurs valeurs de rappels sont obtenus avec l'utilisation de 5 niveaux hiérarchiques. Cela confirme la pertinence de la mesure d'évaluation proposée. en ce qui concerne les résultats de détection, la courbe montre que des améliorations significatives ont été apportées grâce à la structure hiérarchique du VV. En effet, la diversité de l'apparence des régions de tête est atténuée par la flexibilité d'appariement accordée par la structure hiérarchique. Bien que cette flexibilité puisse générer des fausses détections, SVM a été utilisé pour améliorer la précision du système.
Néanmoins, l'algorithme ne parvient pas à détecter un piéton dont la tête est cachée. Ce problème n'est pas observé dans l'ensemble des images de test. Toutefois, le processus de suivi temporel permet de retrouver la nouvelle position du piéton même en cas d'éventuelle occultation de tête.

5 Conclusion

Dans cette communication, nous avons présenté un système original de détection adapté aux problématiques de détection de piétons dans les images d'infrarouge lointain. Le système procède en trois étapes essentielles : détection, validation et suivi, qui se basent toutes sur l'appariement des points d'intérêt SURF extraits dans des régions claires dans l'image. Afin d'atténuer les larges variations de l'apparence des piétons, nous avons crée un Vocabulaire Visuel Hiérarchique. Cette solution apporte plus de flexibilité pour l'étape fondamentale de mise en correspondance. La représentation hiérarchique a permis non seulement d'accélérer la mise en correspondance mais aussi d'améliorer les performances de détection. De plus, les expérimentations montrent que le système proposé produit des résultats précis et robustes face aux problèmes de changements d'échelle et d'occultations partielles.

Références

[1] D. Comaniciu. *Non parametric information fusion for motion estimation*, IEEE Conference on Computer Vision and Pattern Recognition, pp 59-66, 2003.

[2] M. Bertozzi, A. Broggi, A. Fascioli, T. Graf et M.M. Meinecke. *Pedestrian Detection for Driver Assistance Using Multiresolution Infrared Vision*, IEEE Transactions on Vehicular Technology, Vol. 53,pp 1666-1678, 2004.

[3] D.G. Lowe, B. Labbe, A. Rogozan et A. Bensrhair. *Distinctive Image Features from Scale-Invariant Keypoints*, International Journal of Computer Vision, Vol. 60,pp 91-110, 2004.

[4] B. Leibe, E. Seemann et B. Schiele. *Pedestrian detection in crowded scenes*, Computer Vision and Pattern Recognition, pp 20-25, 2005.

[5] H. Bay, T. Tuytelaars, L.J. Van Gool. *SURF : Speeded Up Robust Features*, ECCV, pp 404-417, 2006.

[6] M. Bertozzi, A. Broggi, M. Felisa and G. Vezzoni. *Low-level Pedestrian Detection by means of Visible And Far Infra-red Tetra-vision*, IEEE Intelligent Vehicles Symposium, pp 231-236, 2006.

[7] B. Leibe, A. Ettlin, B. Schiele. *Learning semantic object parts for object categorization*, Image and Vision Computing, Vol. 26, No. 1, pp 15-26, 2008.

[8] K. Jungling et M. Arens. *Feature based person detection beyond the visible spectrum*, Computer Vision and Pattern Recognition Workshop, pp 30-37, 2009.

[9] B. Besbes, B. Labbe, A. Rogozan et A. Bensrhair. *SVM-based fast pedestrian recognition using a hierarchical codebook of local features*, IEEE Workshop on Machine Learning for Signal Processing (MLSP), pp 226-231, 2010.

[10] K. Fukunaga,et L. Hostetler. *The estimation of the gradient of a density function, with applications in pattern recognition*, IEEE Transactions on Information Theory, Vol. 21, No. 1, pp 32-40, 1975.