Détection de bustes et évaluation de descripteurs combinés pour la classification d'orientations

Laurent Fitte-Duval[1,2] Alhayat Ali Mekonnen[1,2] Frédéric Lerasle[1,2]

[1] CNRS, LAAS, 7, Avenue du Colonel Roche, F-31400 Toulouse, France
[2] Université de Toulouse, UPS, LAAS, F-31400 Toulouse, France

{lfittedu,aamekonn,lerasle}@laas.fr

Résumé

Cet article présente des modalités visuelles permettant, dans un contexte d'interaction homme-robot, d'évaluer l'intention d'une personne d'interagir avec un autre agent (homme ou robot). L'analyse de la partie supérieure du corps humain comprenant la tête et les épaules fournit des indices révélateurs de l'intention d'une personne. Nous proposons une approche rapide et efficace pour détecter le buste d'une personne ainsi qu'un classifieur des orientations possibles du buste dans des images 2D. Notre détecteur, dérivé d'un performant détecteur de piétons de la littérature, s'appuie sur les descripteurs ACF [1], obtenus par combinaison de descripteurs hétérogènes, et sur une représentation pyramidale rapide de ces descripteurs durant la détection. Notre classifieur d'orientations utilise une représentation en matrices creuses pour reconnaître l'orientation du buste. Le détecteur présenté montre des résultats comparables à ceux de la littérature sur une base de données publique en termes de précision et de coût CPU. Nous évaluons également différentes combinaisons de descripteurs pour la classification d'orientations présentant des résultats prometteurs malgré les défis associés.

Mots Clef

détection de bustes, classification d'orientations, représentation pyramidale rapide, représentation en matrices creuses, descripteurs ACF, évaluation de descripteurs.

Abstract

This work investigates some visual functionalities required in Human-Robot Interaction (HRI) to evaluate the intention of a person to interact with another agent (robot or human). Analyzing the upper part of the human body which includes the head and the shoulders, we obtain essential cues on the person's intention. We propose a fast and efficient upper body detector and an approach to estimate the upper body pose in 2D images. The upper body detector derived from a state-of-the-art pedestrian detector identifies people using Aggregated Channel Features (ACF) and fast feature pyramid whereas the upper body pose classifier uses a sparse representation technique to recognize their shoulder orientation. The proposed detector exhibits state-of-the-art result on a public dataset in terms of both detection performance and frame rate. We also present an evaluation of different feature set combinations for pose classification using upper body images and report promising results despite the associated challenges.

Keywords

upper body detection, body pose classification, fast feature pyramid, sparse representation, aggregated channel features, feature evaluation.

1 Introduction

En interaction Homme-robot (HRI [2]), l'un des besoins fondamentaux est une détection et une localisation correctes des agents humains au voisinage du robot. Le robot doit être capable de percevoir les actions des agents humains afin de s'y coordonner. Selon l'application, l'espace d'interaction peut varier de quelques centimètres à plusieurs mètres. Cette proximité introduit de nombreuses contraintes sur le champ de vision de la caméra embarquée sur le robot. Habituellement, en interaction proximale, la majorité des caméras embarquées ont une vue partielle des humains, en particulier la partie au-dessus des cuisses (cf. figure 1) ainsi tout mécanisme de détection de personnes adopté doit prendre ce fait en considération.

En HRI, la détection de personnes s'appuie, soit sur des caméras RGB classiques [16], soit sur des caméras RGB-D qui peuvent fournir des données 3D comme le capteur Kinect [13]. En raison de contraintes physiques, économiques et de conception, les caméras RGB classiques sont prédominantes dans les applications robotiques. Ainsi, dans cet article, nous allons nous concentrer sur des perceptions à partir d'images 2D RGB. L'approche la plus populaire pour la détection de personnes en HRI utilise un détecteur de piétons entraîné sur une base de données de personnes annotées [8]. Malheureusement, ces détec-

1. Pour *Aggregated Channel Features*

2. Pour *Human-Robot Interaction*

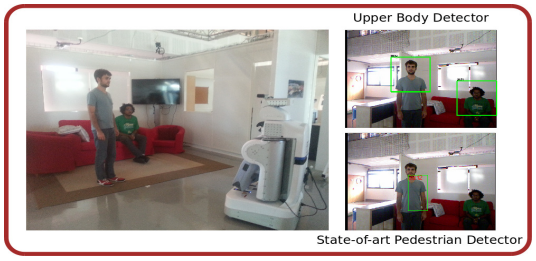


FIGURE 1 – Résultat du détecteur proposé (en haut à droite) et d'un détecteur de piétons de l'état de l'art [7] (en bas à droite) dans un contexte de HRI.

teurs ne parviennent pas à détecter les personnes en présence d'occultations partielles, en particulier l'occultation des jambes. En se concentrant sur la partie supérieure du corps humain, principalement la tête et les épaules, il devient possible d'identifier la présence de l'homme dans une image. Cette approche, désignée comme la détection de bustes, est semblable à une détection de piétons se concentrant sur une zone plus petite moins variable que le corps entier et moins sensible aux occultations [15, 21]. La figure 1 illustre ce point : Pour une situation Homme-Robot (H/R) typique comme celle décrite sur l'image de gauche, le meilleur détecteur de piétons ne parvient pas à détecter correctement les deux personnes face au robot (en bas à droite) alors que notre détecteur de bustes le gère parfaitement (en haut à droite).

Après avoir identifié les agents humains, nous avons besoin de caractériser leur comportement global et leur degré d'intentionnalité d'interagir avec le robot ou avec un autre agent humain. Généralement, l'analyse de ces indices est liée à des indices provenant de l'analyse du visage [14, 17]. Ces indices montrent la direction du regard et indique qui est l'interlocuteur privilégié s'il y a interaction [3].

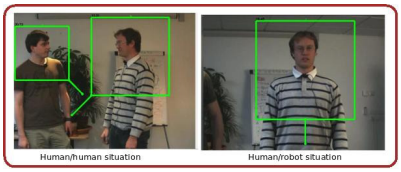


FIGURE 2 – Résultat de notre classifieur en situation d'interaction. Les flèches proéminentes indiquent l'orientation du buste.

Mais l'estimation conjointe des orientations de la tête et du corps permet également de confirmer la position d'un interlocuteur si il y a interaction ou de connaître la direction de mouvement d'une cible mobile [5].

Dans cette optique, nous proposons d'estimer l'orientation de la personne en utilisant des indices provenant du buste. Nous présentons une évaluation approfondie de différentes combinaisons de descripteurs afin d'identifier la meilleure combinaison capturant des indices discriminants nécéssaires pour classifier l'orientation du buste. Pour différents scénarios comme ceux représentés dans la figure 2, l'orientation du buste permet de différencier une situation où deux agents interagissent entre eux sans se préoccuper du robot (à gauche) d'une situation où l'agent fait face au robot en vue d'une éventuelle interaction (à droite).

État de l'art De nombreux chercheurs ont étudié la détection de bustes [15, 21]. Les descripteurs les plus fréquemment utilisés dans ces travaux sont les histogrammes de gradient orienté (HOG) [6] qui capturent la distribution du gradient dans l'image. Ce sont actuellement les descripteurs les plus discriminants en détection et les meilleurs résultats sont obtenus par des approches qui en utilisent des variantes [8]. Certains travaux améliorent les performances de détection en considérant des combinaisons de descripteurs hétérogènes, par exemple, en combinant les motifs binaires locaux (ou descripteur LBP [3]) aux descripteurs HOG [21, 12]. Des progrès récents en détection de personnes mettent l'accent sur la représentation des descripteurs en introduisant la notion de descripteurs ICF [4] [8]. Cette représentation profite du calcul rapide des descripteurs en utilisant les images intégrales et combine des descripteurs hétérogènes pour obtenir des performances de détection dépassant celles des détecteurs de la littérature basés sur le descripteur HOG. Dollár *et al.* (2014) ont proposé une représentation alternative appelée descripteurs ACF qui améliorent légèrement les performances. Ces deux descripteurs ICF et ACF ont obtenu des performances exceptionnelles en termes de détection et de temps de calcul. Combinée à un classifieur en cascade souple [4] et une représentation multi-échelles efficace par approximation des descripteurs redimensionnés durant la détection, cette approche produit le détecteur de piétons le plus rapide de l'état de l'art [7].

L'estimation de l'orientation (ou pose) du buste a aussi été explorée dans la littérature [9, 19]. Ces travaux utilisent différentes méthodes pour récupérer un modèle articulé dans la bonne configuration sur une seule image ou une séquence d'images. Pour éviter la complexité associée à des modèles articulés, certains travaux utilisent les données obtenues à partir de la détection complète des piétons pour estimer leur pose dans un contexte de vidéosurveillance [1, 2, 5]. Ces travaux utilisent des descripteurs HOG calculés à plusieurs échelles avec différents classifieurs similaires à ceux utilisés dans le processus de détection afin d'attribuer une orientation au piéton observé, comme les SVM [5] [1] ou les arbres de décision aléatoires [2]. Chen *et al.* (2012) propose une technique originale de représentation en matrices creuses qui est adoptée dans ces travaux. Cette approche s'est révélée robuste aux occultations et à la corruption des données.

Contributions Cet article décrit deux contributions essentielles : (1) il présente un détecteur de bustes à partir d'images 2D RGB utilisant les descripteurs ACF et un classifieur en cascade souple qui conduit à des résultats probants en termes de performances de détection et de coût CPU ; et (2) il présente une évaluation comparative détaillée de différentes combinaisons de descripteurs pour la

3. Pour *Local Binary Pattern*
4. Pour *Integral Channel Features*
5. Pour *Support Vector Machine*

FIGURE 3 – Synoptique de l'approche de perception adoptée.

classification d'orientations du buste en utilisant une technique de représentation en matrices creuses.

L'article est structuré comme suit : Il commence par un résumé global de notre approche en section 2. Les sections 3 et 4 détaillent les approches utilisées respectivement pour la détection de bustes et la classification d'orientations du buste. Toutes les expérimentations menées et les résultats associés sont présentés dans la section 5. Enfin, le document se termine par une conclusion en section 6.

2 DESCRIPTIF DE NOTRE APPROCHE

La figure 3 illustre synthétiquement l'approche de perception adoptée et ses deux modules de base : le détecteur de bustes (Upper Body Detector) et la classification d'orientations (Body Pose Classification). Comme indiqué dans la section 1, pour une séquence d'images RGB donnée, nous voulons détecter les personnes dans le flux d'images et estimer correctement leur orientation. Le détecteur de bustes repère les personnes dans l'image en utilisant une fenêtre glissante qui scanne exhaustivement l'image à toutes les positions et échelles possibles et en y appliquant le modèle entraîné. Cette étape, très lourde en termes de calcul, est le principal obstacle empêchant d'obtenir des vitesses d'exécution élevées. Aussi, nous privilégions les descripteurs ACF associés à un classifieur en cascade souple qui sont performants en termes de détection et en coût CPU. Ce module fournit ensuite tous les bustes détectés au module de classification.

Le module de classification détermine l'orientation des données de bustes fournis. Pour cela, nous avons adopté une technique basée sur une représentation des données en matrices creuses. Nous évaluons une multitude de combinaisons de descripteurs dont les descripteurs ACF classiques, leurs variantes multi-échelles et une variante multi-niveaux des descripteurs HOG afin de ne conserver que la combinaison de descripteurs menant au meilleur résultat dans notre contexte applicatif. Ce module fournit alors des informations d'orientation pour les bustes détectés.

3 DÉTECTEUR DE BUSTES

Nous présentons les processus clés utilisés dans notre détecteur : la représentation en descripteurs ACF, la structure du classifieur et l'algorithme de représentation multi-échelles des descripteurs.

3.1 Descripteurs ACF

Les descripteurs ACF diffèrent des descripteurs ICF car ils utilisent les valeurs des pixels au lieu de sommes sur des régions rectangulaires [7]. Un canal C est une représentation d'une image I dans laquelle les pixels sont obtenus en appliquant une fonction de génération de descripteurs Ω. Plusieurs canaux $C = \Omega(I)$ peuvent être définis en générant différents types de descripteurs. Les canaux de descripteurs utilisés dans ce travail sont :

La couleur : Les niveaux de gris sont le canal colorimétrique le plus simple à obtenir à partir d'une image car $C = I$. Dans notre application, nous utilisons les trois canaux de couleur LUV qui se sont révélés très efficaces en détection de personnes [8].

L'intensité de gradient (MG [6]) : Une transformation non linéaire qui traduit la force du contour.

Les histogrammes de gradient orienté : Il s'agit d'un histogramme indexé par les six orientations du gradient (une classe par orientation) et pondéré par l'intensité de gradient. La normalisation de l'histogramme par l'intensité de gradient permet une approximation des descripteurs HOG.

Les motifs binaires locaux (ou descripteur LBP) : nous introduisons un nouveau descripteur par rapport à ceux utilisés dans les travaux précédents [7]. Les motifs binaires locaux sont l'un des meilleurs descripteurs de texture de la littérature. Nous en utilisons une implémentation basée sur les niveaux de gris et invariante par rotation inspirée des travaux de Ojala et al., (2002).

Une fois, tous les canaux $C = \Omega(I)$ générés, ils sont divisés en blocs. Les pixels de chaque bloc sont sommés et les canaux de résolution inférieure résultants sont lissés pour obtenir les descripteurs ACF. Ils sont ensuite introduits dans un classifieur boosté exploitant des arbres de décision via une architecture en cascade souple.

3.2 Classifieur en cascade souple

Le boosting est une méthode de classification qui consiste à combiner plusieurs classifieurs faibles pondérés pour créer un classificateur fort. Le classifieur en cascade souple est une variante de boosting proposé par Zhang et Viola (2008) et utilisé pour la détection des piétons avec succès dans [7]. Contrairement à une cascade classique qui présente un nombre prédéfini d'étages distincts, la cascade souple a un seul étage avec de nombreux classifieurs faibles. Les seuils de rejet rendant la structure en cascade "souple" sont calculés pour chaque classifieur faible durant un processus d'étalonnage de la cascade qui permet de réordonnancer les étages "souples" afin d'améliorer la précision de la classification. Son principal atout est la conservation de l'information à chaque étage durant le processus global d'apprentissage contrairement à la cascade attentionelle classique où chaque étage est entraîné séparément. Il permet aussi d'optimiser le taux de détection et le temps d'exécution.

3.3 Représentation pyramidale rapide

Une pyramide de descripteurs est une représentation multi-échelles d'une image où les canaux de descripteurs $C_s = \Omega(I_s)$ sont calculés à chaque échelle s pour l'image redimensionée correspondante I_s. Les échelles sont organi-

6. Pour *Magnitude Gradient*

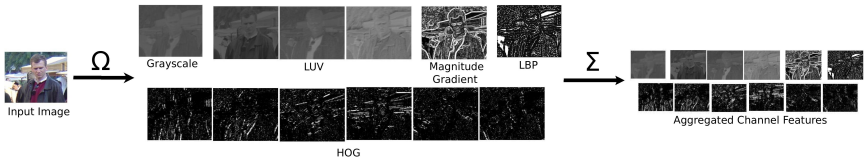


FIGURE 4 – Génération des descripteurs ACF

sées en octaves, des ensembles de 8 échelles uniformément échantillonnées dans l'espace logarithmique entre une échelle donnée et la prochaine échelle correspondant à la moitié de sa valeur. Au lieu de générer chaque échelle $C_s = \Omega(I_s)$, on utilise une approximation du canal de descripteurs redimensionné :

$$C_s = R(C,s) \cdot s^{\lambda_\Omega} \quad (1)$$

où $R(C,s)$ représente le canal C redimensionné à l'échelle s et λ_Ω, un coefficient de puissance lambda spécifique à la transformation Ω afin de redimensionner C. Une méthode itérative peut en être déduite pour générer efficacement une représentation pyramidale des descripteurs en approximant les canaux de descripteurs aux échelles intermédiaires s en utilisant le canal généré à l'échelle la plus proche s' tel que $C_s = R(C, s/s') \cdot (s/s')^{\lambda_\Omega}$. Le meilleur compromis consiste à générer une échelle par octave et d'approximer les 7 autres.

Cette approche, initiée par Dollár et al. (2014), combinée aux descripteurs ACF et au classifieur en cascade souple présentés précédemment permet une mise en œuvre efficace d'une détection par fenêtre glissante. On obtient alors un détecteur rapide et précis comme nous le verrons en section 5.

4 CLASSIFICATION D'ORIENTATIONS DU BUSTE

L'objectif de cette modalité est d'estimer l'orientation du buste d'une personne par rapport à la caméra. La méthode de classification adoptée a déjà été utilisée pour estimer l'orientation de piétons [5].

4.1 Représentation du buste

En considérant l'estimation de l'orientation d'une personne comme un problème de classification - plutôt qu'un problème de régression en raison des difficultés à générer une base de données - et conformément à la littérature, nous considérons huit orientations possibles, espacées également de 45°(N, NE, E, SE , S, SW, W, NW), cf. figure 5. Dans les travaux de [5], des descripteurs HOG multi-niveaux sont générés pour extraire des informations à partir d'une fenêtre englobant une personne. Trois différents niveaux sont générés en utilisant différentes tailles de cellules

FIGURE 5 – Les huit classes d'orientations considérées.

(respectivement 8×8, 16×16 et 32×32) et l'orientation du gradient est quantifiée en 9 classes. Ici, nous proposons d'utiliser les descripteurs ACF pour estimer l'orientation à partir d'une représentation en matrices creuses. Nous profitons également de la représentation pyramidale rapide des descripteurs introduite dans la section 3.3 pour générer des descripteurs ACF multi-échelles à trois niveaux différents similaires à ceux utilisés pour les descripteurs HOG multi-niveaux. Nous générons donc des descripteurs ACF à trois échelles sans approximer les échelles intermédiaires dans l'octave.

4.2 Classification à partir d'une représentation en matrices creuses

La méthode de classification adoptée est la même que celle utilisée dans [5] et inspirée par le travail réalisé en reconnaissance de visages dans [20]. Une matrice d'apprentissage $\mathbf{F} = [F_1, F_2, ..., F_k]$, est générée avec $F_i (1 < i < k) \in \mathbb{R}^{m \times n_i}$, la matrice associée à l'orientation i composée de n_i vecteurs de descripteurs de dimension m. Un nouveau vecteur de descripteurs $\mathbf{y}$ peut être exprimé comme une combinaison linéaire des vecteurs de descripteurs de la base d'apprentissage

$$\mathbf{y} = a_1 F_1 + a_2 F_2 + ... + a_k F_k = \mathbf{F}\mathbf{a}_0 \quad (2)$$

où $\mathbf{F} = [F_1, F_2, ..., F_k]$ et le vecteur $\mathbf{a}_0 = [a_1, a_2, ..., a_k]^T$ est la concaténation des coefficients associés aux vecteurs de chaque classe. Les coefficients de $\mathbf{a}_0$ obtenus en résolvant l'équation $\mathbf{Fa} = \mathbf{y}$ sont des coefficients non nuls si ils sont liés à l'orientation du nouveau vecteur de descripteurs démontrant ainsi l'aspect creux de cette décomposition. Ce problème peut être résolu simplement en réalisant une minimisation l_1 :

$$\mathbf{a*} = \arg\min \|a\|_1 \text{ tel que } \mathbf{Fa} = \mathbf{y}, \quad (3)$$

qui peut être résolue en utilisant la pseudo-inverse de $\mathbf{F}$. Après avoir obtenu cette décomposition, on peut calculer la probabilité de chaque classe d'orientation $\rho_k(\mathbf{y})$ tel que :

$$\rho_k(\mathbf{y}) = \sum a_i^* / \|a*\|_1, \quad (4)$$

Classifier $\mathbf{y}$ consiste alors à trouver la probabilité d'orientation maximale définie par :

$$class(\mathbf{y}) = \max \rho_k(\mathbf{y}) \quad (5)$$

5 EXPÉRIMENTATIONS

5.1 Combinaisons de descripteurs

Les descripteurs ACF sont essentiellement utilisés au sein des deux modules de détection et de classification d'orien-

tations. Utiliser le même type de descripteurs est avantageux car il permet d'éviter des calculs supplémentaires au sein du processus global. Afin d'observer les effets des différents canaux sur le détecteur et le classifieur d'orientations, nous évaluons différentes combinaisons de descripteurs comme énumérées ci-dessous :

— L'intensité de gradient et les six canaux des histogrammes de gradient orienté (GM+HOG ou **a**),
— Le canal couleur, l'intensité de gradient, et les histogrammes de gradient orienté (Clr+GM+HOG ou **b**),
— Le canal couleur, l'intensité de gradient, les histogrammes de gradient orienté et les motif binaires locaux (Clr+GM+HOG+LBP ou **c**),
— L'intensité de gradient, les histogrammes de gradient orienté et les motifs binaires locaux (GM+HOG+LBP ou **d**),

Pour la classification d'orientations, les niveaux de gris sont utilisés à la place des trois canaux de couleur LUV car les bases de données d'apprentissage actuellement disponibles sont principalement composées d'images en niveaux de gris, $(Clr = I)$.

5.2 Bases de données et réglage des paramètres libres

Dans les deux modules, les fenêtres contenant des bustes sont des fenêtres carrées et tronquées à partir de fenêtres englobant des piétons - de même largeur et d'un tiers de la hauteur à partir du bord supérieur (cf. figure 5). La fenêtre de base adoptée est de dimension 64×64. On a observé que l'utilisation d'une fenêtre carrée contenant le tiers d'une fenêtre contenant une personne plutôt que sa moitié conduit à une perte marginale des performances de détection globale sur les bases de données utilisées, mais, en contrepartie, on détecte mieux une personne se trouvant près du robot.

5.2.1 Détection de bustes.

Base de données : Pour entraîner notre détecteur de bustes, nous utilisons la base de données de l'INRIA [6]. La base de données contient 614 images positives contenant 1,237 piétons annotés dans diverses situations notamment des foules. Les échantillons négatifs sont choisis au hasard à partir de 1218 images vides de personnes. A l'instar de [11], la taille de la base d'apprentissage est augmentée en perturbant les échantillons positifs par application de légères rotations (3 rotations de 3^o) et de réflections. La taille du groupe d'échantillons positifs est donc augmentée 6 fois atteignant un total d'environ 7000 échantillons. L'introduction de ces variations permet une meilleure généralisation du classifieur.

Pour tester le détecteur entraîné, nous utilisons les images de test de la base de données InriaLite, un sous-ensemble de la base de données de l'INRIA contenant 145 images de personnes en extérieur montrant un total de 219 personnes, la plupart d'entre eux étant entièrement visibles de face ou de dos.

Apprentissage du détecteur : Le classifieur en cascade souple entraîne 2,048 arbres de décision de profondeur 2 en quatre phases de bootstrapping où les échantillons négatifs mal classés sont réutilisés.

5.2.2 Classification d'orientations du buste.

Base de données : Pour entraîner et tester le classifieur d'orientations, nous avons utilisé la base de données TUD Multiview Pedestrians [1]. Cette base de données contient entre 400 et 749 images de piétons annotées pour chaque classe pour l'apprentissage. Il contient aussi deux bases d'images de validation et de test contenant chacune 248 images annotées, qui sont conjointement utilisées pour réaliser nos évaluations soit un total de 496 échantillons.

Apprentissage du classifieur d'orientations : Comme présenté en section 4.2, la taille de la matrice d'apprentissage $\mathbf{F}$ dépend du nombre d'échantilllons par classe n_i et de la dimension m du vecteur des descripteurs qui varie durant notre évaluation. Les paramètres affectant m sont spécifiques des descripteurs utilisés. Pour les descripteurs ACF, cela dépend du nombre de canaux qui varie entre 7 et 9 dans le module de classification et de la taille des blocs utilisés la phase de compression. Nous utilisons une taille fixe de blocs de 4×4 pixels. Pour être sûr que le classifieur d'orientations ne soit pas surentraîné (en overfit), on teste le classifieur entraîné avec différents groupes d'échantillons dont la taille, n_i, est égale pour chaque classe et varie entre 200 and 400 échantillons de test. Finalement, le classifieur est entraîné en utilisant la totalité de la base d'apprentissage (nombre d'échantillons variable pour chaque classe).

5.3 Métriques d'évaluation

Évaluation des détecteurs : Le protocole d'évaluation des détecteurs utilisé est le même que celui utilisé dans le PASCAL Visual Object Classes Challenge [10]. Le système de détection renvoie une série de fenêtres détectées après analyse de l'image. Ces fenêtres sont obtenues après les phases de détection multi-échelles et de suppression des non-maxima, ce qui évite l'accumulation de fenêtres voisines autour d'une même personne. Pour une fenêtre de détection donnée (BB_{dt}) et une fenêtre de vérité terrain donnée (BB_{gt}), on considère que la détection est correcte si leur recouvrement (overlap) excède 50% :

$$overlap = \frac{BB_{dt} \cap BB_{gt}}{BB_{dt} \cup BB_{gt}} > 0.5 \qquad (6)$$

Les fenêtres BB_{dt} et BB_{gt} non assignées sont respectivement considérées comme des faux positifs et des faux négatifs. La comparaison des détecteurs est réalisée comme dans [8], en traçant la moyenne logarithmique du taux de cibles manquées (MR [7]) en fonction du taux de faux positifs par image (FPPI).

Évaluation du classifieur : La classification est évaluée en utilisant des matrices de confusion où les colonnes corres-

7. Pour *Miss Rate*

pondent aux classes estimées tandis que chaque ligne correspond aux classes de la réalité terrain. Les estimations concentrées le long de la diagonale indiquent une bonne performance. Nous pouvons ensuite moyenner leur précision sur toutes les classes, ce qui constitue notre premier critère de performance, la précision 1 (noté $acc.1$ [8]). On considère un second critère, la précision 2 (noté $acc.2$ [9]), où les estimations sur les 2 classes voisines d'une classe donnée sont considérées comme correctes [1].

5.4 Résultats

Détection de bustes : Pour comparer la performance du détecteur de bustes proposé et ses variantes (utilisant différentes combinaisons de descripteurs) avec la littérature, nous avons utilisé trois autres détecteurs de bustes incluant l'implémentation OpenCV du détecteur de Viola et Jones [18], notre implémentation du détecteur de Dalal et Triggs (2005) basé sur les descripteurs HOG et un classifieur SVM (HOG-SVM), et le détecteur proposé par le groupe Calvin [9] basé sur un modèle par parties (DPM [10]).

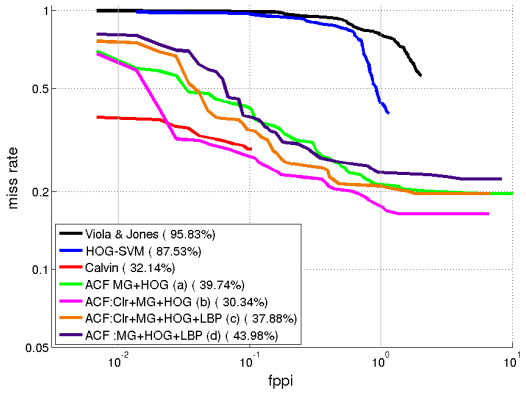


FIGURE 6 – moyenne logarithmique du taux de cibles manquées sur la base InriaLite.

Les résultats de la figure 6 montre que tous les détecteurs utilisant les descripteurs ACF surpassent les détecteurs basés sur les approches de Viola et Jones et HOG-SVM.

Les meilleurs résultats sont obtenus par le détecteur original dont les descripteurs ACF combinent le canal couleur, l'intensité de gradient et les histogrammes de gradient orienté (Clr+GM+HOG) qui surpasse également le détecteur Calvin - le meilleur détecteur de buste de la littérature - atteignant un score global de 30.34% (moyenne logarithmique du taux de cibles manquées). Ce détecteur enregistre un taux de cibles manquées de 0.18 pour une moyenne d'un faux positif par image ce qui peut encore être amélioré en utilisant un mécanisme de filtrage. La combinaison d'intensité de gradient et des histogrammes de gradient orienté est la combinaison de descripteurs utilisée la plus simple, mais elle illustre bien les caractéristiques descriptives de l'orientation du gradient. L'ajout seul des motifs binaires locaux diminue les performances du détecteur tandis que l'utilisation de tous les descripteurs présentés les améliorent relativement mais dans l'ensemble il a tendance à surentraîner le détecteur en raison de l'augmentation du nombre de descripteurs.

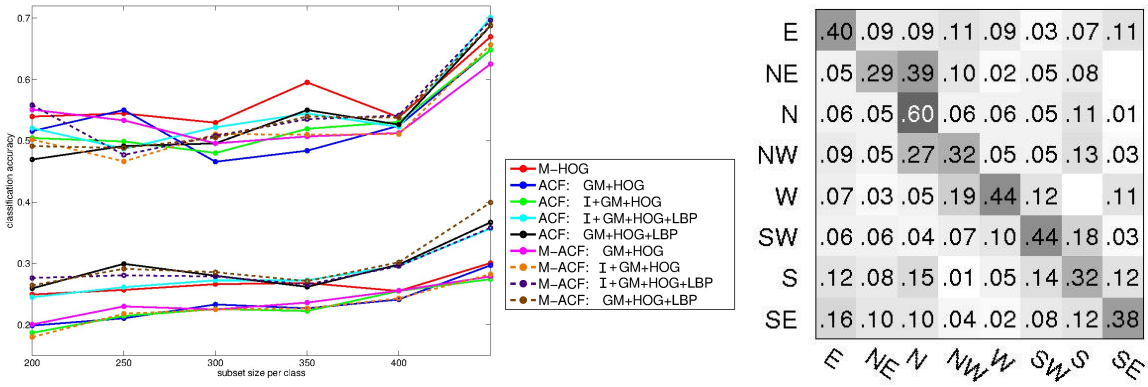
En terme de coût CPU, avec un code Matlab non-optimisé, tous les détecteurs utilisant les descripteurs ACF tournent à environ 12.5 fps [11] pour des images de taille 640×480 sur une machine Intel Core i7 utilisant un seul thread. Cela est suffisant pour les besoins d'une application robotique en temps réel et largement supérieur aux 0.63 images par seconde traitées par le détecteur Calvin.

Classification d'orientations du buste : Le classifieur d'orientations du buste est évalué en utilisant les 496 échantillons. On génère des matrices de classification pour chaque type de descripteurs (11 au total) et on considère les deux critères : $acc.1$ and $acc.2$. On évalue la performance de l'approche proposée et basée sur les descripteurs ACF classiques ainsi que leurs variantes multi-échelles afin de les comparer avec l'approche utilisant les descripteur HOG multi-niveaux [5]. Les résultats correspondants sont présentés dans les figures 7(a) et 7(b). La figure 7(a) montre les variations des performances des classifieurs sur l'ensemble des images de test tandis que le nombre d'échantillons par classe utilisés durant l'apprentissage varie de 200 à 400 échantillons. Les résultats confirment que si l'on utilise davantage de données durant l'apprentissage, il n'y a pas de surapprentissage observé. En effet, les meilleurs résultats sont obtenus en utilisant toutes les données de la base d'apprentissage (correspondant à la dernière série de points des courbes). La matrice de confusion représentée en figure 7(b) correspond au meilleur classifieur obtenu avec la variante multi-échelles des descripteurs ACF combinant intensité de gradient, histogrammes de gradient orienté et motifs binaires locaux (GM+HOG+LBP).

Le tableau 1 récapitule les dimensions et les les précisions obtenues pour la classification d'orientations de buste pour les différentes combinaisons de descripteurs ACF avec ou sans variante multi-échelles, en plus des descripteurs HOG multi-niveaux. Les combinaisons mixant intensité de gradient et histogrammes de gradient orienté montrent des résultats proches de ceux obtenus avec les descripteurs HOG multi-niveaux comme attendu, vu qu'ils utilisent des informations similaires. Mais, contrairement au module de détection, l'amélioration des résultats est obtenue grâce aux informations de texture des motifs binaires locaux tandis que l'addition de l'information couleur en niveaux de gris diminuent les résultats. L'utilisation de l'information multi-échelles permet d'améliorer légèrement la précision puisque la meilleure précision par classe $acc.1$ est obtenue par la variante multi-échelles des descripteurs ACF combinant intensité de gradient, histogrammes de gradient orienté et motifs binaires locaux (GM+HOG+LBP).

La matrice de confusion est très dense avec une concentration des scores sur la diagonale. Certaines erreurs pro-

8. Pour *accuracy 1*
9. Pour *accuracy 2*
10. Pour *Deformable Parts Model*
11. Pour *frames per second*



(a) (b)

FIGURE 7 – (a) Précision de classification d'orientations de buste, *acc*.1 (courbes du bas), *acc*.2 (courbes du haut), tracé comme une fonction de la quantité d'échantillons par classe ; la dernière série de points correspond à une classification en utilisant tous les échantillons de la base de données. (b) Matrice de confusion pour l'estimation d'orientation du buste utilisant la combinaison de descripteurs ACF multi-échelles : GM+HOG+LBP.

viennent de mauvaises classifications pour des classes adjacentes. Par exemple, les orientations Nord-Est et Nord-Ouest sont souvent confondues avec l'orientation Nord. Ces erreurs sont prises en compte par le second critère de précision. Ce critère a le même ordre de grandeur d'environ 66% des estimations quel que soit le type de descripteurs utilisés. Nous pouvons également citer le fait que la meilleure estimation d'orientation est celle indiquant le Nord, correspondant à une personne de dos, moins ambigüe que les autres orientations où le visage est visible. Même les confusions dues à la symétrie sont moins visibles à cause de toutes les faibles estimations pour toute une classe d'orientation. Ces classifications permettent d'avoir un aperçu des interactions dans l'image, mais nécessitent d'être améliorées. Des exemples correctes ou incorrectes de classification d'orientations sont montrés en figure 8.

TABLE 1 – Classification de l'orientation du buste

		upper body (64x64)		
Approche		dim.	acc. 1	acc. 2
HOG multi-niveaux		756	0.3	0.67
Descripteurs ACF	(a)	1792	0.3	0.65
	(b)	2048	0.27	0.65
	(c)	2304	0.36	0.7
	(d)	2048	0.37	0.69
Descripteurs ACF multi-échelles	(a)	2352	0.28	0.63
	(b)	2688	0.28	0.66
	(c)	3024	0.36	0.7
	(d)	2688	0.4	0.7

6 CONCLUSION ET PERSPECTIVES

Dans cet article, nous avons présenté deux importants concepts de perception à partir d'images 2D - la détection de bustes et l'estimation d'orientations du buste - qui ont des applications pertinentes en interaction homme/machine. Le détecteur de bustes développé sur la base de descripteurs ACF présente des résultats probants, en améliorant les performances du précédent meilleur détecteur de 2% en moyenne pour le taux de cibles manquées tandis qu'il améliore de 20 fois son coût CPU. Dans cette optique, nous avons également présenté un classifieur d'orientations du buste basé sur une représentation en matrices creuses qui utilise les descripteurs ACF classiques et multi-échelles. En général, les résultats de la classification d'orientations ont montré une précision comparable à la meilleure approche de la littérature. Les résultats ont été améliorés par l'introduction du descripteur LBP conduisant à une précision de 70% (*acc*.2), en termes de classification d'orientation, surpassant la meilleure approche dans la littérature. Ainsi, les modules de perception proposés sur la base des descripteurs ACF obtiennent de très bonnes performances autant en précision qu'en coût CPU. Dans nos futurs travaux, les fonctionnalités présentées seront couplées avec une approche de filtrage stochastique afin d'améliorer les performances puis intégrées sur nos plates-formes robotiques mobiles pour détecter l'intention de personnes.

ACKNOWLEDGEMENTS

Ces travaux sont financés par les projets ROMEO2 (http ://www.projetromeo.com/) et RIDDLE – BPIFrance, respectivement, dans le cadre des projets structurants des pôles de compétitivité (PSPC), et de la subvention numéro ANR-12-CORD-0003 de l'Agence Nationale de la Recherche française.

Références

[1] Andriluka, M., Roth, S., and Schiele, B., Monocular 3d pose estimation and tracking by detection, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 623-630, 2010.

[2] Baltieri, D., Vezzani, R., and Cucchiara, R., People orientation recognition by mixtures of wrapped dis-

(a) (b)

FIGURE 8 – Quelques illustrations de détections de buste (a), et d'estimations d'orientation (b) provenant, respectivement, des bases de données Inrialite et TUD Multiview.

tributions on random trees,0 *European Conference in Computer Vision (ECCV)*, pp. 270-283, 2012.

[3] Bazzani, L., Cristani, M., Tosato, D., Farenzena, M., Paggetti, G., Menegaz, G., and Murino, V., Social interactions by visual focus of attention in a three-dimensional environment, *Expert Systems*, Vol. 30(2) ,pp. 115-127, 2013.

[4] Bourdev, L. and Brandt, J., Robust object detection via soft cascade, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 236-243, 2005.

[5] Chen, C., Heili, A., and Odobez, J.-M., Combined estimation of location and body pose in surveillance video, *IEEE Advanced Video and Signal-Based Surveillance (AVSS)*, pp. 5-10, 2011.

[6] Dalal, N. and Triggs, B., Histograms of oriented gradients for human detection, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Vol. 1, pp. 886-893, 2005.

[7] Dollár, P., Appel, R., Belongie, S., and Perona, P., Fast feature pyramids for object detection, *Transactions on Pattern Analysis and Machine Intelligence*, Vol. 36(8), pp. 1532-1545 2014.

[8] Dollár, P., Wojek, C., Schiele, B., and Perona, P., Pedestrian detection : An evaluation of the state of the art, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 34(4), pp. 743-761, 2012.

[9] Eichner, M., Marin-Jimenez, M., Zisserman, A., and Ferrari, V., 2D articulated human pose estimation and retrieval in (almost) unconstrained still images, *International Journal of Computer Vision*, Vol. 99(2), pp. 190-214, 2012.

[10] Everingham, M., Van Gool, L., Williams, C. K., Winn, J. and Zisserman, A., The pascal visual object classes (voc) challenge, *International journal of computer vision*, Vol. 88(2), pp. 303-338, 2010.

[11] Ferrari, V., Marin-Jimenez, M., and Zisserman, A., Progressive search space reduction for human pose estimation, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1-8, 2008.

[12] Hu, R., Wang, R., Shan, S., and Chen, X., Robust head-shoulder detection using a two-stage cascade framework, *International Conference on Pattern Recognition (ICPR)*, 2014.

[13] Jafari, O. H., Mitzel, D., and Leibek, B., Real-time RGB-D based people detection and tracking for mobile robots and head-worn cameras, *International Conference on Robotics and Automation (ICRA2014)*, 2014.

[14] Katzenmaier, M., Stiefelhagen, R., and Schultz, T., Identifying the addressee in human-human-robot interactions based on head pose and speech, *International Conference on Multimodal Interfaces (ICMI)*, pp. 144-151, 2014.

[15] Li, M., Zhang, Z., Huang, K., and Tan, T., Rapid and robust human detection and tracking based on omega- shape features, *International Conference on Image Processing (ICIP)*, pp. 2545-2548, 2009.

[16] Mekonnen, A. A., Lerasle, F., and Zuriarrain, I., Multi-modal person detection and tracking from a mobile robot in a crowded environment, *International Conference on Computer Vision Theory and Applications (VISAPP 2011)*, pp. 511-520, 2011.

[17] Sheikhi, S. and Odobez, J.-M., Recognizing the visual focus of attention for human robot interaction, *Human Behavior Understanding*, pp. 99-112, Springer, 2012.

[18] Viola, P. and Jones, M., Rapid object detection using a boosted cascade of simple features, *IEEE Conference on Computer Vision and Pattern Recognition*, Vol. 1, pp. I-511, 2001.

[19] Weinrich, C., Vollmer, C., and Gross, H.-M.,Estimation of human upper body orientation for mobile robotics using an svm decision tree on monocular images, *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 2147-2152, 2012.

[20] Wright, J., Yang, A. Y., Ganesh, A., Sastry, S. S., and Ma, Y., Robust face recognition via sparse representation, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 31(2), pp. 210-227, 2009.

[21] Zeng, C. and Ma, H., Robust head-shoulder detection by pca-based multilevel HOG-LBP detector for people counting, *International Conference on Pattern Recognition (ICPR)*, pp. 2069-2072, 2010.