Extraction de silhouettes à partir de boîtes englobantes de détection

Cyril MEURIE[1], Olivier LÉZORAY[2], Marion BERBINEAU[1]

[1]Univ Lille Nord de France, F-59000 Lille, IFSTTAR, COSYS, LEOST, F-59650, Villeneuve d'Ascq
[2]Normandie Univ, UNICAEN, ENSICAEN, CNRS, GREYC, 14000 Caen, France
{cyril.meurie,marion.berbineau}@ifsttar.fr,olivier.lezoray@unicaen.fr

Résumé – Dans de nombreuses applications telles que la vidéosurveillance, la détection de personnes est un élément clé. Habituellement, une boîte englobante contenant une personne est extraite. Cependant, une segmentation plus précise de la silhouette peut s'avérer nécessaire. Pour ce faire, nous proposons une stratégie complète de segmentation des personnes à l'intérieur de boîtes de détection. Elle s'appuie sur plusieurs étapes : pré-traitement, localisation approximative de la silhouette des personnes et raffinement de la silhouette à l'aide de coupes de graphes. Comme de nombreuses méthodes différentes (et leurs paramètres associés) peuvent être choisies pour chaque étape, nous utilisons une approche génétique pour déterminer la stratégie de segmentation optimale et montrons ses avantages par rapport à l'état de l'art.

Abstract – In many applications such as video surveillance, people detection is a key element. Usually, a bounding-box containing a person is extracted. However a more precise segmentation of peoples' silhouette can be needed. To that aim, we propose a complete scheme for people segmentation inside detection bounding boxes. This relies on several steps: pre-processing, approximate people silhouette localisation and silhouette refinement with graphs cuts. Since many different methods (and associated parameters) can be chosen fo each step, we use a genetic approach towards determining the optimal segmentation scheme and show its benefit towards the state-of-the-art.

1 Introduction

La détection des personnes est un élément clé des systèmes de vidéosurveillance. C'est un problème difficile en raison des multiples configurations possibles qui peuvent se produire [1]. De nombreuses approches ont été proposées en utilisant des caractéristiques dédiées [2, 3] qui sont fournies à des méthodes d'apprentissage telles que des Support Vector Machines. Les méthodes de détection des personnes fournissent habituellement un résultat sous la forme d'une boîte englobante autour de la personne qui a été détectée. Cependant, ceci n'est pas toujours suffisant pour de nombreuses applications évoluées telles que la reconnaissance des personnes : la silhouette de la personne à l'intérieur de la boîte englobante est nécessaire. Peu de travaux ont abordé la segmentation de la silhouette des personnes [4, 5]. Nous proposons une nouvelle stratégie de segmentation des silhouettes de personnes à partir de boîtes englobantes de détection. Pour ce faire, nous considérons plusieurs pré-traitements d'images associés à différentes méthodes de segmentation basées sur différentes caractéristiques. Ces méthodes fournissent plusieurs cartes de probabilités qui sont combinées puis raffinées par une coupe de graphe. Chaque méthode implique néanmoins des paramètres qui doivent être réglés manuellement. Pour éviter cela, nous proposons de déterminer automatiquement avec un algorithme génétique la meilleure stratégie de segmentation dans son ensemble (choix des méthodes constituant la stratégie), avec les meilleurs paramètres des méthodes, ce qui permet en outre un temps de traitement réduit (voir [6] pour plus de détails). La stratégie est éprouvée sur des bases de référence et comparée à l'état de l'art.

2 Synopsis de l'approche

La finalité est d'extraire les silhouettes de personnes dans des boîtes englobantes obtenues par une précédente phase de détection. Nous considérons huit stratégies de segmentation différentes mais qui reposent toutes sur l'enchainement de trois étapes communes (Figure 1). ① Pré-traitement : plusieurs types

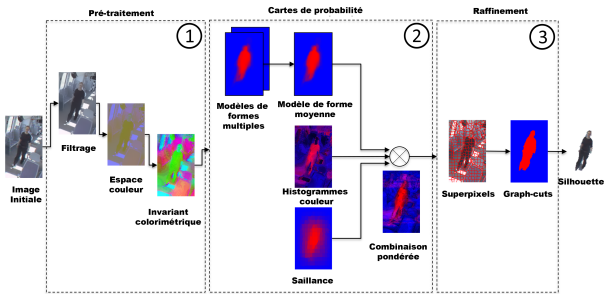

FIGURE 1 – Synopsis de notre approche.

de pré-traitements préalables sont considérés. ② Estimation d'une carte de probabilité : estime la probabilité qu'un pixel appartienne à l'arrière-plan ou à la silhouette. Différents a priori d'apparences peuvent être considérés et combinés. ③ Raffinement : une coupe de graphe classe les pixels en deux classes à partir de la carte de probabilité. Ceci peut être fait au niveau pixel ou superpixel. Les huit stratégies sont obtenues en variant les techniques utilisées dans chaque étape. Cette conception modulaire a été retenue afin de pouvoir évaluer l'intérêt de chaque étape et maîtriser la complexité et la vitesse de calcul de l'ensemble de la stratégie.

3 Stratégies proposées

Avant de détailler le contenu de chaque stratégie, nous détaillons les différentes méthodes qui peuvent être considérées pour chaque étape.

3.1 Pré-traitement

Trois méthodes consécutives de pré-traitement peuvent être effectuées : un changement d'espace couleur (RGB, HSL, YUV, $L^*a^*b^*$, $L^*u^*v^*$), un filtrage (Gaussien, médian, bilatéral) et un invariant couleur (Greyworld, Reduced coordinates, $l_1l_2l_3$, $m_1m_2m_3$, affine normalization, RGB rank). Chacune de ces méthodes peut permettre à l'étape de pré-traitement de réduire le bruit, de mieux différencier certaines couleurs, et d'être plus robuste à l'illumination.

3.2 Cartes de probabilité

La deuxième étape permet de déterminer une probabilité d'appartenance au premier plan (silhouette des personnes) ou à l'arrière plan pour chaque pixel. Ces informations sont extraites de trois différents types de méthodes : les modèles de formes, les histogrammes de couleurs et la saillance. Trois modèles de formes et deux types d'histogrammes couleur sont considérés. Lorsque plusieurs cartes de probabilités sont extraites par plusieurs méthodes, elles sont combinées avec une combinaison pondérée.

3.2.1 Modèles de forme

L'utilisation de modèles de forme est courant dans la littérature pour estimer la position d'une personne dans une boîte englobante [7, 8].

Modèle de forme moyenne Comme nous utilisons des boîtes englobantes résultats d'un processus de détection de personne, celles-ci sont généralement centrées sur la personne. Comme dans [5], on peut alors utiliser un modèle de forme moyen. Celui-ci est calculé une fois pour toutes et est obtenu à partir d'une moyenne de toutes les vérité terrain de silhouettes d'une base d'apprentissage.

Modèle de formes multiples Un simple modèle de forme moyen ne peut pas bien rendre compte des différentes poses et angles de vue qui peuvent se produire. Pour y pallier, nous considérons de multiples modèles de formes et choisissons automatiquement le meilleur [5] afin de disposer d'un modèle de forme adapté à la posture de la personne. La vérité terrain des silhouettes est découpée en k groupes (par un k-moyennes) ce qui fournit de multiples modèles de formes. Ceci est fait au préalable et k est optimisé par un algorithme génétique dédié. Ensuite deux méthodes sont considérées pour choisir le modèle de forme de silhouette le plus plausible pour une boîte englobante. Une première méthode considère des histogrammes de couleurs d'arrière-plan et de silhouette et le modèle de forme maximisant la différence entre les deux histogrammes est retenu. Une seconde méthode considère des histogrammes de gradients orientés (HOG) fournis à un SVM qui détermine le meilleur modèle de silhouette, après un apprentissage préalable en un-contre-tous. Les paramètres du HOG et du SVM sont également déterminés au préalable par un algorithme génétique dédié.

3.2.2 Histogrammes couleur

Le deuxième type de méthode considéré repose sur l'apparence couleur des pixels à partir de deux types d'histogrammes construits pour l'arrière-plan et la silhouette. Ces histogrammes sont construits relativement au modèle de forme choisi automatiquement. Le premier type d'histogramme est défini sur toute la boîte englobante alors que le second est le résultat de la concaténation d'histogrammes locaux extraits par bandes. Les appartenances des pixels sont ensuite estimées à partir des distributions modélisées par les histogrammes.

3.2.3 Saillance

Étant donné que l'objet le plus saillant à l'intérieur de la boîte englobante est censé être la personne qui s'y trouve, les méthodes de détection d'objets saillants peuvent être considérées comme de bonnes candidates pour l'estimation de la probabilité d'arrière-plan ou de premier-plan. Nous avons considéré pour cela la méthode de [9].

3.2.4 Combinaison de cartes de probabilité

Les trois types de méthodes présentés précédemment fournissent 6 cartes de probabilités différentes d'appartenance au fond $P_k^{back}(p_i)$ ou bien à la silhouette $P_k^{front}(p_i)$ pour un pixel p_i et une méthode k. Si une méthode a des paramètres, ils seront fixés de manière optimale avec un algorithme génétique dédié. Nous les combinons en une seule carte à l'aide d'une combinaison pondérée : $P^{cl}(p_i) = \sum_k \theta_k P_k^{cl}(p_i)$ avec $\sum_k \theta_k = 1$ et cl parmi $back$ ou $front$. Une optimisation génétique sera utilisée pour déterminer les coefficients optimaux θ_k.

3.3 Raffinement

À partir de la carte de probabilité finale obtenue par combinaison, l'étape finale consiste à estimer les classes des pixels. Nous exploitons pour cela une coupe dans un graphe. Ce graphe $G = (V, E)$ peut être un graphe des pixels ou bien de superpixels obtenus avec [10]. L'objectif est alors d'assigner un label $l_i \in L = \{0, 1\}$ aux noeuds $p_i \in V$. L'énergie suivante est considérée et minimisée avec le min-cut/max-flow de [11] :

$$\hat{\imath} = \underset{l \in F}{\operatorname{argmin}} \left(\sum_{p_i \in V} W^{l_i}(p_i) + t \sum_{p_i \in V} \sum_{p_j \in N_{p_i}} S(p_i, p_j) \cdot \delta_{l_i \neq l_j} \right) \quad (1)$$

La meilleure segmentation (en deux classes : silhouette et arrière-plan) correspond au minimum de l'énergie $\hat{\imath}$ dans l'ensemble F de toutes les solutions d'étiquetage possibles. Le premier terme est $W^{l_i}(p_i) = -\gamma * \log(P^{l_i}(p_i))$ et il exploite les probabilités initiales estimées. Le terme $\delta_{l_i \neq l_j}$ encourage un étiquetage

constant par morceaux, N_{p_i} est l'ensemble des voisins de p_i avec les autres sommets du graphe. Le terme $S(p_i, p_j)$ est une similarité entre p_i et p_j donné par :

$$S(p_i, p_j) = \alpha * exp\left(-\frac{d(p_i, p_j)}{2 * \beta^2}\right) \cdot \frac{1}{dist(p_i, p_j)} \qquad (2)$$

avec $dist(p_i, p_j)$ une distance spatiale et $d(p_i, p_j)$ une distance colorimétrique (effectuée sur les moyennes dans le cas de superpixels). Les paramètres α, β, γ ont une forte influence sur le résultat final et ils seront déterminés immédiatement par l'optimisation génétique globale.

4 Optimisation génétique des stratégies

Lorsque nous choisissons une configuration spécifique pour chacune des trois étapes de la stratégie, nous obtenons des stratégies de segmentation spécifiques, mais différents. L'ensemble des stratégies retenues est résumé dans le Tableau 1.

TABLE 1 – Méthodes considérées pour chaque stratégie.

		Stratégies proposées (nombre de gènes entre parenthèses)							
Méthode	**[5]**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
① **Pré-traitement**									
Espace Couleur		✓(1)	✓(1)	✓(1)	✓(1)	✓(1)	✓(1)	✓(1)	✓(1)
Filtrage						✓(4)			
Invariant Couleur						✓(1)			
② **Carte de Probabilité**									
Modèle de forme multiples									
Difference d'histogramme								✓(3)	
HOG + SVM									✓(1)
Modèle de forme moyenne	✓	✓(0)	✓(0)	✓(0)	✓(0)	✓(0)	✓(0)		
Histogrammes couleur			✓(2)						
Bandes d'histogrammes				✓(3)	✓(3)	✓(3)	✓(3)	✓(3)	✓(3)
Saillance					✓(3)				
Combination pondérée			✓(2)	✓(2)	✓(3)	✓(2)	✓(2)	✓(2)	✓(2)
③ **Raffinement**									
Superpixels							✓(3)		
Graph-cuts	✓	✓(3)	✓(3)	✓(3)	✓(3)	✓(3)	✓(3)	✓(3)	✓(3)
Nombre total de genes		(4)	(8)	(9)	(13)	(14)	(12)	(12)	(10)

Ces stratégies ont chacune des paramètres associés qui influent sur chacune des étapes. Plutôt que de les régler manuellement, nous proposons de le faire par optimisation génétique afin de déterminer la meilleure configuration d'une stratégie donnée. Une population de chromosomes encode les solutions possibles à explorer. Chaque chromosome correspond à l'encodage de l'ensemble des méthodes et paramètres d'une stratégie. Un chromosome est divisé en plusieurs blocs qui sont composés d'un ou plusieurs gènes. Chaque gène code soit l'utilisation d'une méthode (gène binaire) soit la valeur d'un paramètre (valeurs possibles quantifiées). Par exemple le bloc correspondant à l'utilisation d'une méthode de filtrage est composé de 4 gènes : un gène est utilisé pour le choix du filtre et les trois autres sont utilisés pour les paramètres du filtre. L'algorithme génétique utilise une configuration standard et se compose de quatre étapes (initialisation, sélection, croisement et mutation) jusqu'à stabilisation. Comme nous utilisons un encodage spécifique de la solution en blocs dans les chromosomes, les opérations de croisement et mutation sont spécifiques (Figure 2). Une mutation peut concerner soit le gène qui encode le choix d'une méthode et dans ce cas les gènes associés au paramètres mutent également, ou bien seulement le gène qui encode une valeur de paramètre. Pour le croisement la recopie des gènes se fait par blocs. Ceci est illustré dans la Figure 2 pour la partie du chromosome qui encode l'étape de pré-traitement. Les huit stratégies retenues dans le Tableau 1 sont les huit meilleures solutions de la population finale.

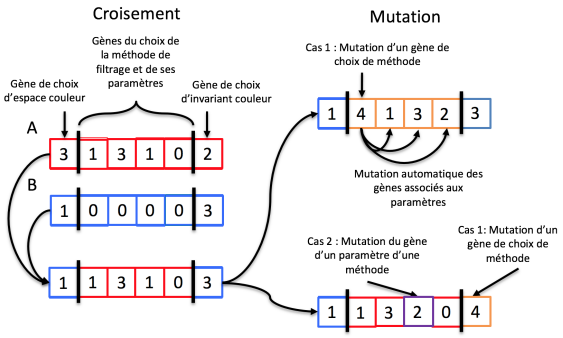

FIGURE 2 – Exemple de croisement et mutation de notre algorithme génétique.

5 Résultats

Pour tester les stratégies de segmentation proposées, nous considérons plusieurs bases de données provenant de tâches de détection et de reconnaissance de personnes, et pour lesquelles les images fournies proviennent d'une méthode basée sur la détection des personnes sous la forme d'une boîte englobante (éventuellement de taille différentes pour une même base). Nous avons également sélectionné des bases de données permettant d'avoir des poses, des angles de vue et des problèmes d'éclairage variés. Au total nous avons considéré 6 bases de données, avec un total de 1797 images. Ce sont les bases ViPeR [1], people re-ID 2011 [12] (PRID 2011), INRIA Person dataset [13], et BOSS [14]. Pour les bases qui ne fournissaient pas de segmentation de référence des silhouettes, nous les avons créées. Pour chaque stratégie de segmentation, l'ensemble d'apprentissage est utilisé par l'algorithme génétique et la F-mesure est utilisée comme mesure de fitness. Pour plus d'indépendances vis-à-vis des données d'apprentissage, nous effectuons une 8-fold validation croisée. Avec une telle validation croisée, on obtient une estimation moyenne de la performance d'une stratégie, mais cela en génère 8 différentes. Nos retenons uniquement celle de plus grande F-mesure. La stratégie est ensuite évaluée sur les données de test (à nouveau par 8-fold validation croisée). Le tableau 2 montre les résultats obtenus pour chaque base de données avec chaque stratégie de segmentation. Les résultats sont divisés en deux catégories : F-mesure et temps de traitement. Chaque stratégie a été testée sur un thread CPU cadencé à 3.4Ghz. Pour une comparaison équitable, la stratégie de référence de l'état de l'art de [5] a également été optimisée par notre algorithme génétique. Les résultats apparaissant en vert sont équivalents ou supérieurs à [5] et ceux apparaissant en rouge correspondent à la meilleure stratégie. Nous mettons en gras le résultat de la stratégie qui nous semble la meilleure en terme de compromis efficacité/rapidité. Les résultats de [5] sont bons sur toutes les bases, même si la méthode est très simple et rapide. Les stratégies 1 à 3 ajoutent à [5] un changement d'espace couleur, la stratégie

TABLE 2 – F-mesure et temps de traitement pour les stratégies de segmentation proposées et l'état de l'art.

		Stratégies proposées							
Dataset	**[5]**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
BOSS S-1	**0.899**	0.897	0.905	**0.913**	0.912	0.916	0.840	0.900	0.895
(160x96) [14]	**16ms**	16ms	22ms	**23ms**	118ms	25ms	110ms	39ms	28ms
BOSS S-2	**0.857**	0.862	0.871	**0.883**	0.881	0.881	0.810	0.860	0.851
(160x96) [14]	**16ms**	16ms	25ms	**25ms**	121ms	203ms	108ms	33ms	27ms
INRIA	**0.839**	0.837	0.854	**0.859**	0.860	0.859	0.790	0.845	0.832
(128x64) [13]	**9ms**	11ms	14ms	**15ms**	73ms	60ms	82ms	18ms	15ms
INRIA	**0.839**	0.839	0.852	**0.855**	0.850	0.856	0.800	0.848	0.835
(160x96) [13]	**15ms**	15ms	24ms	**25ms**	114ms	110ms	110ms	32ms	29ms
VIPeR	**0.887**	0.887	0.887	**0.894**	0.900	0.894	0.830	0.883	0.864
(128x48) [1]	**10ms**	10ms	12ms	**13ms**	86ms	82ms	74ms	17ms	15ms
PRID2011	**0.818**	0.832	0.894	**0.900**	0.900	0.876	0.813	0.882	0.822
(128x64) [12]	**12ms**	12ms	18ms	**20ms**	84ms	38ms	81ms	33ms	29ms
Moyenne	**0.856**	0.859	0.877	**0.884**	0.883	0.880	0.814	0.870	0.850
	13ms	13ms	19ms	**20ms**	99ms	86ms	94ms	26ms	24ms

2 un modèle d'apparence à base d'histogrammes globaux, la stratégie 3 un modèle d'apparence à base d'histogrammes en bandes et cela permet des améliorations. La stratégie 3 fournit toujours des résultats bien meilleurs que les résultats de base de [5] et cela montre les avantages des étapes considérées. Cela entraine cependant un léger surcoût de calcul, mais qui reste compatible avec du temps-réel. L'ajout d'une information de saillance (stratégie 4) peut permettre d'améliorer les résultats sur quelques bases, mais au détriment du temps de calcul. L'ajout de pré-traitement de filtrage et d'invariants couleur permet également des gains, mais à nouveau au détriment du temps de calcul. La stratégie 6 reprend la stratégie 3 avec des superpixels pour l'étape de raffinement, et cela entraine une perte de précision due au fait que la décision est prise au niveau du superpixel et celui-ci peut être imprécis. Enfin les stratégies 7 et 8 remplacent le modèle de forme moyen de la stratégie 3 par des modèles de formes multiples. Le temps de traitement reste comparable à la stratégie 3, mais les résultats sont un peu en deçà. Ce résultat peut sembler surprenant, mais confirme des résultats similaires obtenus dans [7, 5]. La stratégie 3 est au final la stratégie que nous retenons pour le bon compromis qu'elle réalise entre performances et temps de traitement tout en surpassant l'état de l'art de [5]. La figure 3 présente des résultats sur PRID2011.

FIGURE 3 – Extraction de silhouettes sur PRID2011[12] (images originales, résultats de [5] et de la stratégie 3).

6 Conclusion

Nous avons considéré l'extraction de silhouettes dans des boîtes englobantes et proposé une stratégie qui va au-delà de l'état de l'art. Elle comprend un changement d'espace couleur, une combinaison pondérée d'a priori d'apparences et un raffinement par coupe de graphe. Une optimisation des méthodes constituant la stratégie a été réalisée par un algorithme génétique.

Références

[1] S. Gong, M. Cristanio, S. Yan, and C. Loy, *Person re-identification*. Springer, 2014.

[2] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *CVPR*, 2005, pp. 886–893.

[3] Q. Zhu, S. Avidan, M.-C. Yeh, and K.-T. Cheng, "Fast human detection using a cascade of histograms of oriented gradients," in *CVPR*, 2006, pp. 1491–1498.

[4] N. Jojic, A. Perina, M. Cristani, V. Murino, and B. Frey, "Stel component analysis : Modeling spatial correlations in image class structure," in *CVPR*, 2009, pp. 2044–2051.

[5] C. Migniot, P. Bertolino, and J-M.Chassery, "Automatic people segmentation with a template-driven graph cut," in *ICIP*, 2011, pp. 3149–3152.

[6] C. Coniglio, C. Meurie, O. Lézoray, and M. Berbineau, "People silhouette extraction from people detection bounding boxes in images," *Pattern Recognition Letters*, vol. 93, pp. 182–191, 2017.

[7] C. Migniot, P. Bertolino, and J.-M. Chassery, "Contour segment analysis for human silhouette pre-segmentation," in *VISAPP*, vol. 2, 2010, pp. 74–80.

[8] Z. Lin and L. S. Davis, "Shape-based human detection and segmentation via hierarchical part-template matching." *IEEE T PATTERN ANAL*, pp. 604–618, 2010.

[9] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang, "Saliency detection via graph-based manifold ranking," in *CVPR*, 2013, pp. 3166–3173.

[10] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE T PATTERN ANAL*, vol. 34, no. 11, pp. 2274–2282, 2012.

[11] Y. Boykov and V. Kolmogorov, "An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision," *IEEE T PATTERN ANAL*, vol. 26, no. 9, pp. 1124–1137, 2004.

[12] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof, "Person re-identification by descriptive and discriminative classification," in *Image Analysis*. Springer, 2011, pp. 91–102.

[13] N. Dalal and B. Triggs, "Inria person dataset," INRIA, Tech. Rep., 2005.

[14] Boss, "Boss europeen project (on board wireless secured video surveillance)," 2009. [Online]. Available : https ://www.celticplus.eu/project-boss/